博士論文

背景雑音と話者の違いに頑健な音声認識

2013年1月28日

指導教員　峯松 信明 教授

電気系工学専攻

37-107091　鈴木 雅之

あらまし

音声認識は様々なシステムの要素技術として利用されている．例えば，カーナビシステム，スマートフォンの音声対話システム，企業のコールセンタにおける電話自動応対システムなど，その応用範囲は多岐に渡る．音声認識の精度を高めることは，これらのシステムのユーザ満足度を向上させることに直結する．そのため，音声認識の精度を向上させるために研究を進めていくことが重要である．

音声認識の精度は，様々な要因によって低下してしまうことが知られている．例えば背景雑音が音声に重畳してしまった場合，何も対処を行わないと音声認識精度は大幅に低下してしまう．他にも，話者の違い，マイクとの距離，部屋の残響，話している内容と，様々な要因によって音声認識精度が低下してしまう．

本論文では，背景雑音と話者の違いに対して頑健なシステムを構築することで，より精度の高い音声認識の実現を目指す．音声認識に関するこれまでの研究サーベイの結果，本論文で特に注目したのは，特徴量ドメインでの雑音抑圧と，識別的リランキングにおける音声の構造的表象の利用である．特徴量ドメインでの雑音抑圧により音声認識が背景雑音に頑健な，識別的リランキングにおける音声の構造的表象の利用により話者の違いに頑健な音声認識を実現することを目指す．

本論文では，まず 第 1 章 で音声認識技術の基本について述べた後，第 2 章 において現在の state-of-the-art 音声認識システムで利用されている重要な技術をサーベイし，どの点に注目して研究を進めることが音声認識システム全体の精度向上につながるのか考察する．その考察に基づき，第 3 章 ではクリーン音声状態の識別に基づく特徴量ドメインでの雑音抑圧手法を，第 4 章 ではクリーン音声に対しても精度低下のない雑音抑圧手法を，第 5 章では識別的リランキングにおける音声の構造的表象の利用を提案し，その有効性を示す．最後に 第 6 章 で本論文をまとめ，今後の課題と展望について述べる．

目次

図目次

表目次

第1章

音声認識の基本技術

1.1　音声認識研究の背景

音声認識は様々なシステムの要素技術として利用されている．一般ユーザ向けとしては，例えば以下のシステムが挙げられる．

- カーナビシステム
- スマートフォンの音声対話システム
- スマートフォンのテキスト入力システム
- 外国語の対話支援システム
- 障がい者支援システム

一般ユーザ向けの音声認識は，キーボードが使えない，もしくは使いにくい状況における，テキストやコマンド入力機能として用いられることが多い．例えば運転中で手がふさがっており，カーナビを手で操作することが難しい．またスマートフォンも，キーボードが比較的打ちづらい状況である．このような場合，音声認識をキーボードやボタンの代替手段として使う必然性がある．また，外国語での対話や，聴覚障がい者の会話など，そもそも音声によるコミュニケーションが難しい場合のコミュニケーション支援システムなどにも，音声認識が利用されている．

またビジネス分野では，例えば以下のようなシステムに音声認識が使われている．

- コールセンターの電話自動対応システム
- 議事録自動作成システム
- 電子カルテの入力システム
- コールセンターにおける電話の録音音声からのマイニング

ビジネス分野では，人員コストを削減するソリューションとして，音声認識が使われることが多い．例えばコールセンターの電話自動対応システムや議事録作成は，音声認識で（半）自動化することで，直接的に人員コストを削減することが可能になる．またキーボードを利用するよりも音声認識の方が効率の高い特殊な状況下，例えば電子カルテの入力などにおいても，音声認識が用いられることがある．他にも，コールセンターにおける大量の電話の録音音声の書き起こしを作成し，そこから業務改善のための情報をマイニングしたり，企業コンプライアンス違反のチェックするといった応用にも，音声認識が利用できる．

以上で述べたように音声認識には幅広い応用分野がある．音声認識分野の国内の市場規模（推定年間販売額）は 2007 年調べで 60 億円，世界だと 2006 年調べで 10 億ドルを越えると言われている．これらの額は，2012 年現在ではさらに増加していると予想されている．世界の音声認識市場規模 10 億ドルのうち，内訳はコールセンター向けシステムが 6 億ドル，電子カルテなど医療分野向けシステムが 1.7 億ドル，カーナビや携帯電話などのエンドユーザ向けシステムは 1.25 億ドルである．携帯電話向けのシステムは，スマートフォン

の普及により，2006 年以降，市場規模が特に増加していると予想できる．比較として，画像を使った顔認識分野の世界の市場規模は 2008 年調べで 1.86 億ドル，テキストマイニング分野の日本の市場規模は 2010 年調べで 24.6 億円である．一概に結論付けることはできないが，メディア情報処理技術の中で，音声認識は比較的大きな市場があると言える．

特にコールセンター向けの音声認識の市場規模が大きいために，多くの日本のエンドユーザの感覚より音声認識が広く用いられていることを強調しておきたい．日本の音声認識エンドユーザからは，カーナビやスマートフォンで音声認識を利用するケースが多いためか，「カーナビでもスマホでも音声認識は声を出すのが恥ずかしいし，誤りがあって使いづらい．結局手で入力した方が楽で早い．音声認識は使えない技術だ」といった意見を聞くことが多い．しかし実際には，ビジネス向けの音声認識ソリューションの需要が大きいため，音声認識はビジネス的に使える技術と言える．

1.2　音声認識の基本構成

音声認識の目標は，音声信号からテキストを推定することである．本節ではまず，音声認識の基本的な構成を説明する．本節の内容は，音声認識の基本についてであるので，音声認識に関する教科書を読めばさらに詳しい内容を見つけることが出来る [1][2] [1]．

1.2.1　特徴量抽出

まず，マイクで録音された音声信号から，ヒトの発声機構からくる音声の特性や，ヒトの聴覚の特性を考慮しながら，テキスト情報を推定しやすい MFCC（Mel Frequency Cepstrum Coefficients）と呼ばれる特徴量を抽出する（図 1.1）．まず音声信号は，0 から 8 kHz 程度の帯域にテキスト情報が多く含まれることから，16 kHz サンプリング程度で録音する．次にヒトの聴覚は，周波数領域の高域により敏感であることから，高域強調フィルタをかける．この処理はプリエンファシスとも呼ばれる．高域強調フィルタとしては，$1 - 0.97z^{-1}$ などが用いられる．

次に，ヒトの発声機構の動きの速さから考えて，生成された音声信号は 25 ミリ秒程度の範囲内で定常であると仮定できることから，25 ミリ秒程度の時間フレームを持つ窓関数（例えばハミング窓）を，10 ミリ秒程度時間をずらしながらかけていく．そして，各時間窓の音声信号に対し短時間フーリエ変換を行い短時間スペクトルを得る．フーリエ変換に相当する周波数解析は，ヒトも耳の奥にある蝸牛と呼ばれる器官で行なわれている．また，ヒトが音声信号の位相情報に比較的鈍感ということから，短時間フーリエ変換により得られたスペクトルを二乗しパワースペクトルにすることで位相情報を消し，ヒトの音声の大

[1]音声認識の基礎技術を一から勉強する場合は，毎年 8 月に ALAGIN が開催している「音声認識・音声対話技術講習会」に参加することも一手である．

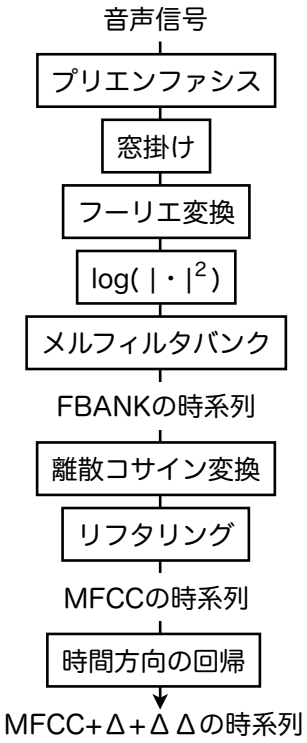

図 1.1: 音声信号からの MFCC+Δ+ΔΔ の時系列の抽出

きさに関する感覚尺度はパワーの対数でおよそ出来ることから，パワースペクトルの対数をとって対数パワースペクトルを得る．

次に，0 から 8 kHz の帯域におけるヒトの聴覚の周波数解像度を考慮して，ヒトの聴覚の音の高低の間隔尺度を近似するメル尺度上で等間隔に並ぶ 20 程度のフィルタバンクをかける．本論文では，この結果得られる特徴量を FBANK と呼ぶ．さらに，FBANK の周波数軸方向に緩やかな成分が，よりテキストを推定するのに有効な情報が含まれていることが知られていることから，FBANK に離散コサイン変換をかけ，そのうちの低次元 13 次元程度のみを残し，そしてそれを定数倍してスケールをおよそそろえる．この処理はリフタリングと呼ばれる．この結果得られる特徴量を MFCC と呼ぶ．

さらに，音声が時間的にどう変化したかの情報にも，テキスト情報が含まれているため，前後 2 つ程度ずつのフレームの MFCC を，時間方向に回帰する．これを及び Δ 特徴量と呼ぶ．さらに，Δ の Δ 特徴量も特徴量として有効であることが知られていおり，これを ΔΔ 特徴量と呼ぶ．また Δ と ΔΔ を含む，時間方向成分を考慮した特徴量を，動的特徴量と呼ぶ．最終的に音声信号は，39 次元程度の MFCC+Δ+ΔΔ が時間方向にならんだ時系列特徴量に変形される．

最近では MFCC の代わりに，PLP（Perceptual Linear Prediction）及びその Δ, ΔΔ が用いられることもある．PLP の方が MFCC より雑音に強いと言われており，実際，IBM,

Microsoft，Google などの音声認識システムでは PLP が採用されている．

PLP の抽出法を 図1.2 に示す．PLP の抽出処理は，MFCC の抽出処理と対応関係が取

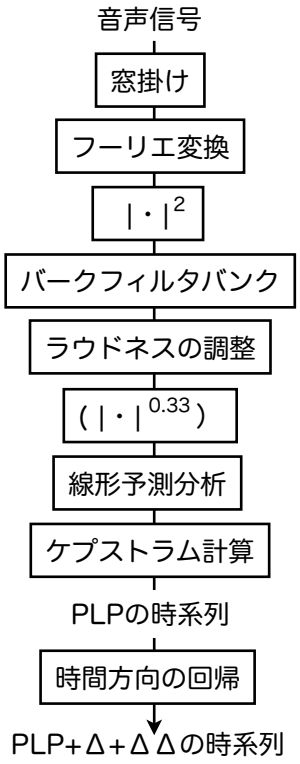

図 1.2: 音声信号からの PLP+Δ+ΔΔ の時系列の抽出

れる部分が多い．まず，バークフィルタバンクは，メルフィルタバンクとよく似た，ヒトの聴覚の周波数解像度を考慮したフィルタバンクである．ラウドネスの調整とは，音の高さの違いによるヒトの聴覚感覚の違いを調整するもので，MFCC におけるプリエンファシスの高域強調と似た処理になっている．0.33 乗をとるのは，スティーヴンスのべき法則に従った処理で，ラウドネスをヒトの間隔尺度であるインテンシティーに変換することに相当する．この処理は，MFCC では該当する処理は存在しない．線形予測分析及びケプストラムの計算は，MFCC における対数をとって DCT 及びリフタリングをする処理に相当している．

MFCC や PLP の抽出の実装は，HTK（`http://htk.eng.cam.ac.uk/`）の HCopy コマンドが利用するのが簡単である．また，Matlab における MFCC 抽出の実装として，rastamat（`http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/`）も利用できる．なお MFCC や PLP を抽出する際に対数演算がある都合上，音声波形が 0 埋めされていると MFCC や PLP が計算できなくなることに注意されたい．このような場合，極微小のホワイトノイズを付加することで問題が回避できる．また SPTK（`http:`

//sp-tk.sourceforge.net/）に含まれているコマンド「mfcc」は，上記で説明した MFCC とは異なる定義が使われているので混同しないようにも注意されたい．

1.2.2　音声認識の定式化

音声信号を特徴量の時系列化した特徴量を，$\boldsymbol{X} = \{\boldsymbol{x}_t\}_{t=1\cdots T}$ とおく．ここで t は時間窓に対応するフレームのインデックスであり，$\boldsymbol{x}_t$ は時刻 t における，例えば 39 次元程度の MFCC+Δ+ΔΔ ベクトル，T はフレームの総数である．我々の目的は，$\boldsymbol{X}$ から対応するテキスト，すなわち単語系列 $\hat{W} = \{\hat{w}_i\}$ を推定することである．例えば $\hat{W} = \{$私 は 鈴木 です$\}$ であれば，$\hat{w}_1 =$ 私, $\hat{w}_2 =$ は, $\hat{w}_3 =$ 鈴木, $\hat{w}_4 =$ です，となる．

この問題を統計的な枠組みで解くことを考えると，音声認識は $\mathrm{argmax}_W\, p(W|\boldsymbol{X})$ を求める問題として定式化できる．これは以下のように変形できる．

$$\begin{aligned}
\hat{W} = \underset{W}{\mathrm{argmax}}\, p(W|\boldsymbol{X}) &= \underset{W}{\mathrm{argmax}}\, \frac{p(\boldsymbol{X}|W)p(W)}{p(\boldsymbol{X})} && (1.1)\\
&= \underset{W}{\mathrm{argmax}}\, p(\boldsymbol{X}|W)p(W) && (1.2)\\
&= \underset{W}{\mathrm{argmax}}\, \log p(\boldsymbol{X}|W) + \log p(W) && (1.3)
\end{aligned}$$

ここで，$\log p(\boldsymbol{X}|W)$ を計算するためのモデルを音響モデル，$\log p(W)$ を計算するためのモデルを言語モデルと呼ぶ．

1.2.3　音響モデル

音響モデルには，left-to-right 型の隠れマルコフモデル（Hidden Markov Model; HMM）が，各状態において混合ガウス分布（Gaussian Mixture Model; GMM）に従って特徴量（MFCC+Δ+ΔΔ）を出力するモデルを用いることが多い．HMM/GMM のモデルパラメタは，HMM の状態の遷移確率，GMM の各正規分布の重み・平均・分散等であり，これはバウムウェルチアルゴリズムを使って最尤推定することができる．計算量の観点から，GMM の各正規分布の共分散行列は対角行列と仮定されることが多い．

w_i のとりうる値が少ない場合（例えば数字しか発声されないと分かっているような場合）は，単語ごとに適当な状態数の HMM/GMM を学習する．

もし w_i のとりうる値が多い場合（大語彙音声認識の場合）は，音素ごとに，3 状態程度の HMM/GMM を学習しておく．認識時には，まず W を対応する音素系列に変換する．例えば $W = \{$私 は 鈴木 です$\}$ の例では，対応する音素系列は {sil w a t a sh i w a s u z u k i d e s u sil} となる（sil は無音の意）．例えば，/suzu/ という部分のみを取り出すと，HMM/GMM は 図1.3 のようになっている．HMM/GMM とビタビアルゴリズムを利用することで，$\log p(\boldsymbol{X}|W)$ が計算できる．

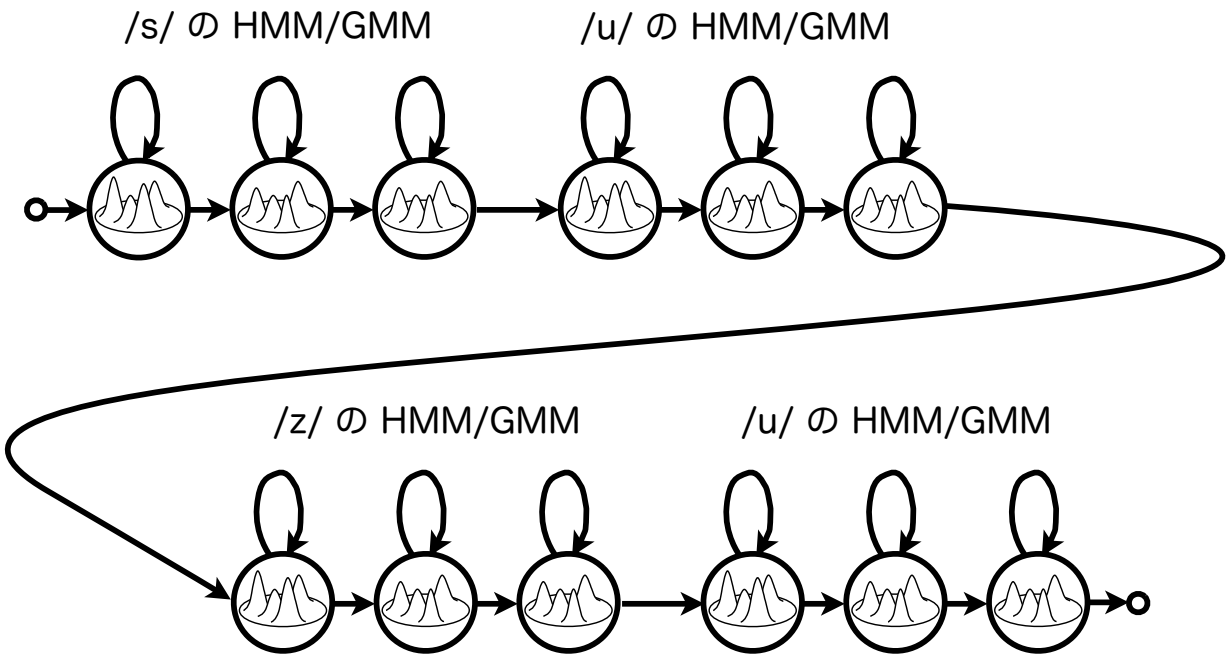

図 1.3: /suzu/という単語が与えられた時に観測特徴量が出力される確率をモデル化した HMM/GMM

なお音素単位の HMM/GMM を使う場合には，観測される $\boldsymbol{x}_t$ は，前後の音素にも影響をうけるため，実際には前後の音素コンテキストを考慮した単位の HMM/GMM を用いることが多い．前後 1 音素ずつのみを見る場合には，この単位は triphone と呼ばれ，$W = \{$ 私 は 鈴木 です $\}$ の例では {sil sil-w+a w-a+t a-t+a t-a+sh a-sh+i sh-i+w i-w+a w-a+s a-s+u s-u+z u-z+u z-u+k u-k+i k-i+d i-d+e d-e+s e-s+u s-u+sil sil} となる．前後 2 音素ずつ合計 5 音素を見る，pentaphone が使われる場合もある．

Triphone や pentaphone を使うと，その数が音素の数の 3/5 乗と大きくなりすぎて，学習データが不足する問題が生じるので，似ている triphone/pentaphone 同士のパラメタを共有して学習を行う．どの triphone/pentaphone のパラメタを共有するかの決定には，決定木アルゴリズムが広く利用されている．

音響モデル作成の実装には，特徴量抽出にも利用した HTK（`http://htk.eng.cam.ac.uk/`）のコマンド群が利用できる．

1.2.4　言語モデル

言語モデルは，単純な少数語彙の単語音声認識であれば，すべての単語が等確率で出現するようなモデルを使えばよい．また，話される内容が非常に定形的な場合であれば，文法を人手で書くことも可能である．

大語彙音声認識の言語モデルには，単語 N-gram を用いることが一般的で，学習データ

の規模に合わせて，N は 2,3,4 などが用いられる．$N = 3$ の場合，w_i がある単語となる確率が，w_{i-2}, w_{i-1} のみに依存するというモデルである．すなわち

$$\begin{aligned} p(W) &= p(w_1, w_2, \cdots w_I) && (1.4) \\ &= p(w_1)p(w_2|w_1)p(w_3|w_2, w_1)\cdots p(w_i|w_{i-1}, w_{i-2})\cdots p(w_I|w_{I-1}, w_{I-2}) && (1.5) \end{aligned}$$

とモデル化する．このモデルは簡単すぎるモデルに思えるが，計算コストや実用性の面から，現在でも広く利用されている．$p(w_x|w_y, w_z)$ の学習は，学習データにおいて w_x の出現した回数を f_x，w_y, w_z の次に w_x が出現した回数を $f_{y,z\to x}$ と置くと，

$$p(w_x|w_y, w_z) = \frac{f_{y,z\to x}}{f_x} \tag{1.6}$$

で最尤推定できる．

単語 3-gram モデルではすべての取りうる w_x, w_y, w_z の組み合わせについて $p(w_x|w_y, w_z)$ を保持しなければならないが，大語彙の音声認識だと，数が多くなりすぎて，学習データに現れないような組み合わせも存在してくる．単語 4-gram などを使うと，この問題はより顕著になる．そこで，学習データに合われなかった場合でも $p(w_x|w_y, w_z)$ が適切に計算できるように，スムージングを行うことが一般的である．スムージングアルゴリズムとしては，以下で示す modified Kneser-Ney smoothing 最も高い性能を示す方法として知られている [3].

$$\begin{cases} p(w_x|w_y, w_z) = \dfrac{f_{y,z\to x} - D(f_{y,z\to x})}{f_x} & \text{if } f_{y,z\to x} > 0 \\ p(w_x|w_y, w_z) = \dfrac{f_{w_y,w_z}}{f_{w_y}}\beta_{w_y,w_z} & \text{if } f_{y,z\to x} = 0 \end{cases} \tag{1.7}$$

ただし，f_{w_y,w_z}, f_{w_y} は，それぞれ，学習データに含まれる $[*, w_y, w_z]$，$[*, w_y, *]$ の総数である．また β_{w_y,w_z} は，$\sum_{w_x} p(w_x|w_y, w_z) = 1$ と確率の定義を満たすように決定するパラメタである．また $D(f_{y,z\to x})$ は以下のように定義される数である．

$$\begin{cases} D(f_{y,z\to x}) = 1 - 2\dfrac{N(1)}{N(1)+2N(2)}\dfrac{N(2)}{N(2)} & \text{if } f_{y,z\to x} = 1 \\ D(f_{y,z\to x}) = 2 - 3\dfrac{N(1)}{N(1)+2N(2)}\dfrac{N(3)}{N(2)} & \text{if } f_{y,z\to x} = 2 \\ D(f_{y,z\to x}) = 3 - 4\dfrac{N(1)}{N(1)+2N(2)}\dfrac{N(4)}{N(3)} & \text{if } f_{y,z\to x} \geq 3 \end{cases} \tag{1.8}$$

ここで $N(i)$ は，学習データの中で i 回現れた n-gram の総数を表す．

N-gram による言語モデルの実装には，例えば SRILM（`http://www.speech.sri.com/projects/srilm/`）や MITLM（`http://code.google.com/p/mitlm/`）などのツールキッ

トが利用できる．

1.2.5　デコーディング

以上の HMM/GMM による音響モデル，N-gram による言語モデルが得られれば，任意の $\boldsymbol{X}$ と W に対して $\log p(\boldsymbol{X}|W) + \log p(W)$ を計算することが可能になる．ここで経験的に，$\mathrm{argmax}_W \log p(\boldsymbol{X}|W) + \log p(W)$ とするよりも，I を W に含まれる単語の数，α と β を適当な定数として，$\mathrm{argmax}_W \log p(\boldsymbol{X}|W) + \alpha \log p(W) + \beta I$ としたほうが音声認識の精度が向上することが知られている．ここで α は音響モデルと言語モデルのどちらに重きを置くのかを調整するパラメタで，β は単語の個数が増えすぎることを防ぐ挿入ペナルティを調整するパラメタである．α と β は，バリデーション用データを使ってグリッドサーチするか，もしくはより簡易的に手動で適当に設定される．

$\mathrm{argmax}_W \log p(\boldsymbol{X}|W) + \alpha \log p(W) + CI$ を得るためには，取りうるすべての W のうち，$\log p(\boldsymbol{X}|W) + \alpha \log p(W) + CI$ が最も大きくなるような W を選ばなければならず，W が多数の単語からなっていたり，N-gram の N が大きくなると，計算量が爆発してしまう．そこで，様々な制限を導入したデコーディングアルゴリズムを用いて適切に探索・枝狩りを行う．デコーディングアルゴリズムを工夫することで，はじめて現実的な計算時間内で妥当な認識結果を得ることができようになる．

デコーディングのアルゴリズムとしては近年，音声認識の探索ネットワークを WFST（Weighted Finite State Transducer）で表現してから探索を行う手法が広く利用されている．WFST を利用したデコーダとしては，例えば Juicer（`http://juicer.amiproject.org/juicer/`）や，Kaidi（`http://kaldi.sourceforge.net/about.html`）に含まれているデコーダなどがある．また日本では，WFST を利用していない，2 パス型の Julius（`http://julius.sourceforge.jp/`）も，オープンソースのデコーダとして有名である．デコーダの改良は計算時間や精度に直結するため，音声認識を扱っている各社は，それぞれ社内専用の性能の良いデコーダを実装しているようである．

第2章

近年の音声認識技術とその問題点

2.1　はじめに

本論文では，音声認識精度を着実に高めることで，音声認識業界全体にインパクトを与えることを目指す．そのために本章では，近年の音声認識技術を深く広く調査し，どういった点を改良するのが研究価値が高いのか考察する．考察の結果本論文では，特徴量ドメインにおけるステレオデータに基づく雑音抑圧と，識別的リランキングにおける多用な音響特徴量の利用の二つを，既存の音声認識の改良すべきポイントと考える．そしてこれに対し本論文において，クリーン音声状態の識別に基づく高速で非定常雑音に頑健な特徴量強調手法，ミスマッチがない場合にも頑健な特徴量強調法，識別的リランキングにおける音声の構造的表象の利用を提案する．

提案法に関してはそれぞれ第 3 章，第 4 章，第 5 章で詳しい説明を行う．これに先立ち本章では，広く有効性が認められている既存の音声認識の要素技術を調査し，なぜこれらのポイントに注目すべきなのかを説明する．

2.2　近年の音声認識技術

第 1 章で説明した技術を基本として，音声認識の精度をさらに向上させるために，様々な手法が研究されている．本節では，それらの手法の中でも，現在でも広く利用されていたり，注目を集めているような要素技術を紹介する．

最新の音声認識技術全般に関する話題は，IEEE Signal Processing magazine の 2012 年 Volume: 29, Issue: 6 の “Fundamental Technologies in Modern Speech Recognition” という特集号が非常に詳しい（`http://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=6296521&punumber=79`）．音声認識全体の話題に関しては，特に [4] で述べられているので，適宜参照されたい．

2.2.1　音声区間検出

マイクで収録される音声には，音声をはなしていない無音の区間が含まれているそれを音声認識するためには，前処理として，音声区間を切り出すことが必要になる．このタスクを音声区間検出（Voice Activity Detection; VAD）と呼ぶ．

簡単には，パワーやゼロ交差数を見ることで VAD を行う．VAD は，実用上音声認識の精度に大きな影響を与えるので，カルマンフィルタを用いる手法など，さまざまな研究が進められている [5].

2.2.2　音響モデルの話者適応

音声認識の精度を高めるための重要なポイントの一つは，モデルの学習に用いるデータと評価に用いるデータにミスマッチが少なくするようにすることである．ミスマッチが存在する場合には，精度が大幅に低下してしまうことが知られている．

ミスマッチの大きな原因の一つに，話者の違いがある．それぞれの話者は，それぞれ固有の形の声道の形状を持っているため，話者によって音声の音色が異なり，MFCC+Δ+ΔΔ も異なる．この話者の違いによるミスマッチを低減するために広く用いられる手法として，音響モデルの HMM/GMM の話者適応がある．

i)　MLLR

HMM/GMM の話者適応技術として有名なのが，MLLR（Maximum Likelihood Linear Regression）適応法である．MLLR 適応は，HTK にも実装されており，非常に広く使われている話者適応技術である．

用意するものは，沢山の話者の音声データから学習した不特定話者音響モデル（HMM/GMM）と，少量の認識対象話者の音声とその正解ラベル（単語系列）である．HMM/GMM の GMM の各要素の正規分布の平均ベクトルすべてに，単一の線形変換をかけることで，ミスマッチを低減させる．そのような線形変換は，上記のデータを用いれば，尤度を最大化するように推定することが出来る．

また，認識対象話者とその正解ラベルデータがたくさんある場合には，HMM/GMM に含まれる沢山の正規分布を予めクラスタリングしておき，それぞれのクラスタごとに異なる線形変換をかけるように MLLR 適応を行えば，さらにミスマッチを低減することが出来る．

MLLR の学習データとして正解ラベルが得られない場合にも，認識対象話者の音声のみさえあれば，その音声の音声認識結果を話者適応用の正解ラベルとして用いることで，教師なし MLLR 適応を実現することが出来る．教師なし MLLR 適応を行う際には，音声認識が間違っている場合に間違った方向に適応が行われてしまうので，適切にクロスバリデーションを行うとよい [6].

2.2.3　特徴量の話者適応

音響モデルの話者適応は，音響モデルを特徴量に合うように変換を施す技術であったが，逆に，特徴量を音響モデルに合うように変換することも可能である．

i)　VTLN

特徴量の話者適応技術しては，声道長正規化（Vocal Tract Length Normalization; VTLN）が古くから知られている．声道形状が，形状を保持したまま大きくなったり小さくなったりするのは，周波数軸でのウォーピング操作に相当する．このウォーピング操作は，声道長を表すパラメタがあれば，それを補償するような変換が実現できる．声道長パラメタの調整は，MLLR 適応と同様，少量の対象話者音声と正解データを用いるか，教師なしで対象話者音声を音声認識した結果を正解データとして，最も尤度が高くなるように調整する．ここで調整すべき声道長パラメタはスカラー値なので，線形変換を推定する MLLR 適応より過学習が起こる可能性が低い．そのため教師なしで求めても十分な精度が実現出来るので，教師なしでパラメタを求めるケースが多いようである．

ii)　fMLLR

さらに，Constrained MLLR （CMLLR）適応も，特徴量の話者適応技術として有名である [7]．CMLLR は，もともとは音響モデルの話者適応技術として提案された技術で，通常の MLLR 適応では GMM の平均ベクトルのみに対して線形変換をかけていたのに対し，平均と分散の数学的な関係を考慮して 一つの線形変換で平均と分散を同時に適応法である．これは，特徴量に対して単一の線形変換をかけることと数学的に同値になっており，CMLLR は，feature-space MLLR（fMLLR）とも呼ばれている．fMLLR と MLLR は，それぞれ単独でも効果があるが，両方利用することで，単独の手法よりも高い精度が得られることが知られている．

2.2.4　音響モデルの雑音適応

雑音環境の違いによっても，モデルと入力のミスマッチが発生するので，このミスマッチを低減してやる必要がある．特に，音声信号と雑音の比（SN 比）が 20dB 以下になると，著しく認識精度が低下することが知られている．

まず，入力の雑音環境で録音された音声及びその正解ラベルがあれば，話者適応で利用した MLLR 適応とまったく同じ枠組みで雑音適応を実現できる．

また，波形領域において観測される信号は，音声波形と雑音波形の単純な足し算になっている関係を利用すると，ある雑音環境で録音された音声がなくても，雑音の推定値さえ得られれば，音響モデルの雑音適応を実現することができる．ここで観測する音声には，音声が含まれていない無音部分があり，その部分は雑音のみが観測されると考えられるため，雑音はは比較的容易に推定することができる．ここで雑音環境は，話者の違いと比べて，時間と共に大きく変動する場合が多いので，雑音のみから音響モデル適応を実現することは重要な技術といえる [8].

クリーンな音声の特徴量を $\boldsymbol{X} = \{\boldsymbol{x}_t\}_{t=1\cdots T}$，雑音が重畳したノイジーな音声の特徴量を $\boldsymbol{Y} = \{\boldsymbol{y}_t\}_{t=1\cdots T}$ とする．また，重畳した雑音の特徴量 $\boldsymbol{n}$ が，時間によらず $p(\boldsymbol{n}) = \mathcal{N}(\boldsymbol{n}; \boldsymbol{\mu}^{\boldsymbol{n}}, \boldsymbol{\sigma}^{\boldsymbol{n}})$ なる対角共分散を持つ正規分布から出力されていると仮定する．そのパラメタである $\boldsymbol{\mu}^{\boldsymbol{n}}, \boldsymbol{\sigma}^{\boldsymbol{n}}$ は，例えば観測音声の最初と最後の無音区間（雑音のみの区間）の特徴量の平均・分散を計算することで単純に求めてもよいし，雑音や音声に対する仮定（雑音音声と比べてゆっくり変化する，時間周波数平面上で音声はスパースになっている，など）を利用して効果的に推定することもできる [9].

ここで，$\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{n}$ の関係について述べる．説明の簡単のために，特徴量として FBANK を用いることにする[1]．ノイジーな音声は，波形領域でクリーン音声と雑音の足し算となっていることを考えると，位相を無視すれば，以下の近似式が成立する．

$$\boldsymbol{y}_t \approx g(\boldsymbol{x}_t, \boldsymbol{n}) = \log\left(\exp \boldsymbol{x}_t + \exp \boldsymbol{n}\right) \tag{2.1}$$

ただし，ここで $\log, \exp$ は，このように，観測されるノイジーな音声特徴量は，ベクトルの各要素ごとの log 及び exp を返す関数とする．すなわち，ノイジーな音声特徴量は，クリーン音声特徴量と雑音特徴量の非線形関数として近似される．

我々が得たいものは，$p(\boldsymbol{Y}|W)$ を計算するための，ノイジーな音声特徴量の音響モデル HMM/GMM である．我々既にクリーンな音声特徴量の音響モデルとして HMM/GMM を持っている．そこで，$p(\boldsymbol{Y}|W)$ を以下のように変形する．

$$\begin{aligned} p(\boldsymbol{Y}|W) &= \prod_t p(\boldsymbol{y}_t|k_t) && (2.2)\\ &= \prod_t \int p(\boldsymbol{x}_t, \boldsymbol{y}_t|k_t) d\boldsymbol{x}_t && (2.3)\\ &= \prod_t \int p(\boldsymbol{y}_t|\boldsymbol{x}_t, k_t) p(\boldsymbol{x}_t|k_t) d\boldsymbol{x}_t && (2.4) \end{aligned}$$

ただし，(2.2) の k_t は，W を使ってデコーディングすることで得られる，時間 t における HMM/GMM の各コンポーネントの正規分布のインデックスである．(2.4) の積分の第二項に，$p(\boldsymbol{x}_t|k_t)$ があり，これは既に持っているクリーン音声特徴量の音響モデルが正規分布としてモデル化しており，計算することが可能である．

[1] MFCC を用いる場合は，これに FBANK から MFCC に変換する線形変換・その逆変換を適切にかければよい．また，Δ や $\Delta\Delta$ を考慮することも可能である．

$p(\boldsymbol{y}_t|\boldsymbol{x}_t, k_t)$ は，さらに以下のように展開できる．

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t, k_t) = \int p(\boldsymbol{y}_t, \boldsymbol{n}|\boldsymbol{x}_t, k_t)d\boldsymbol{n} \tag{2.5}$$

$$= \int p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{n}, k_t)p(\boldsymbol{n})d\boldsymbol{n} \tag{2.6}$$

$$= \int p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{n}, k_t)\mathcal{N}(\boldsymbol{n}; \boldsymbol{\mu}^{\boldsymbol{n}}, \boldsymbol{\sigma}^{\boldsymbol{n}})d\boldsymbol{n} \tag{2.7}$$

(2.5) から (2.6) では，クリーン音声特徴量と雑音特徴量に相関関係がないことを仮定している．

i)　PMC

PMC（Parallel Model Combination）と呼ばれる手法では，(2.2) の $p(\boldsymbol{y}_t|k_t)$ が正規分布に従うと仮定して，(2.7) を (2.4) に代入したものから，モーメント法やサンプリングなどを行うことによって，その正規分布のパラメタを推定する [10]．ただし，モーメント法では精度が悪く，サンプリング法では計算時間がかかりすぎる問題があり，近年では次に紹介する VTS 適応が用いられるケースの方が多い．

ii)　VTS 適応

VTS（Vector Taylor Series）適応では，(2.1) を VTS 近似することで計算を行う [11]．まず，$\boldsymbol{y}_t$ が (2.1) の VTS 近似の値となる場合に，$p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{n}, k_t) = 1$ となり，それ以外では 0 となると近似することで，(2.7) を $\boldsymbol{y}$ の正規分布で近似する．音響モデルの HMM/GMM の GMM の各コンポーネントの正規分布のインデックス k_t ごとに VTS 近似の値を変えることができるので，$\boldsymbol{x}_t$ に関する VTS 近似ではその平均ベクトルを VTS 近似の中心にする．また $\boldsymbol{n}$ の VTS 近似では雑音モデルの平均ベクトルを VTS 近似の中心とする．

1 次の $\boldsymbol{n}$ の VTS 近似を行うと，(2.7) 積分が解析的に解ける．また 1 次の $\boldsymbol{x}_t$ の VTS 近似を行えば，$p(\boldsymbol{y}_t|\boldsymbol{x}_t, k_t)$ が $\boldsymbol{x}_t$ 正規分布の線形変換となる．これを (2.4) に代入すれば，$p(\boldsymbol{x}_t|k_t)$ も $\boldsymbol{x}_t$ 正規分布なので (2.4) の積分が解析的に解け，最終的にノイジーな音声特徴量の音響モデルを得ることができる．以上の理由のため，1 次の VTS 近似が広く利用されている．

$\boldsymbol{y}_t \approx g(\boldsymbol{x}_t, \boldsymbol{n}, k_t)$ を，それぞれ $\boldsymbol{\mu}^{\boldsymbol{x}}_{k_t}$，$\boldsymbol{\mu}^{\boldsymbol{n}}$ を中心として $\boldsymbol{x}_t$，$\boldsymbol{n}$ を 1 次の VTS 近似すると，

以下のようになる．

$$\begin{aligned}
\boldsymbol{y}_t &\approx g(\boldsymbol{x}_t, \boldsymbol{n}, k_t) = \log\left(\exp \boldsymbol{x}_t + \exp \boldsymbol{n}\right) && (2.8)\\
&\approx g(\boldsymbol{\mu}^{\boldsymbol{x}}_{k_t}, \boldsymbol{\mu}^{\boldsymbol{n}}) + \left(\frac{\partial g(\boldsymbol{\mu}^{\boldsymbol{x}}_{k_t}, \boldsymbol{\mu}^{\boldsymbol{n}})}{\partial \boldsymbol{x}_t}\right)^{\top} (\boldsymbol{x}_t - \boldsymbol{\mu}^{\boldsymbol{x}}_{k_t}) + \left(\frac{\partial g(\boldsymbol{\mu}^{\boldsymbol{x}}_{k_t}, \boldsymbol{\mu}^{\boldsymbol{n}})}{\partial \boldsymbol{n}}\right)^{\top} (\boldsymbol{n} - \boldsymbol{\mu}^{\boldsymbol{n}}) && (2.9)\\
&= g(\boldsymbol{\mu}^{\boldsymbol{x}}_{k_t}, \boldsymbol{\mu}^{\boldsymbol{n}}) + \frac{\boldsymbol{x}_t - \boldsymbol{\mu}^{\boldsymbol{x}}_{k_t}}{\mathbf{1} + \exp(\boldsymbol{\mu}^{\boldsymbol{n}} - \boldsymbol{\mu}^{\boldsymbol{x}}_{k_t})} + \frac{\boldsymbol{n} - \boldsymbol{\mu}^{\boldsymbol{n}}}{\exp(\boldsymbol{\mu}^{\boldsymbol{x}}_{k_t} - \boldsymbol{\mu}^{\boldsymbol{n}}) + \mathbf{1}} && (2.10)
\end{aligned}$$

$\log, \exp$ 及び分数演算はそれぞれ log，exp，掛け算，割り算をベクトルの要素ごとに行う演算とし，$\mathbf{1}$ は 1 を並べたベクトルとする．このとき，$p(\boldsymbol{y}_t|\boldsymbol{x}_t, k_t)$ が (2.10) が成り立つときに 1，それ以外で 0 と仮定して (2.7) に代入することで，$p(\boldsymbol{y}_t|\boldsymbol{x}_t, k_t)$ が以下のように計算できる．

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t, k_t) = \mathcal{N}\left(\boldsymbol{y}_t; g(\boldsymbol{\mu}^{\boldsymbol{x}}_{k_t}, \boldsymbol{\mu}^{\boldsymbol{n}}) + \frac{\boldsymbol{x}_{k_t} - \boldsymbol{\mu}^{\boldsymbol{x}}_{k_t}}{\mathbf{1} + \exp(\boldsymbol{\mu}^{\boldsymbol{n}} - \boldsymbol{\mu}^{\boldsymbol{x}}_{k_t})}, \frac{\boldsymbol{\sigma}^{\boldsymbol{n}}}{(\exp(\boldsymbol{\mu}^{\boldsymbol{x}}_{k_t} - \boldsymbol{\mu}^{\boldsymbol{n}}) + \mathbf{1})^2}\right) \quad (2.11)$$

ただし，exp，分数，二乗は，すべてベクトルの要素ごとの計算とする．さらに，これを (2.4) に代入すれば，

$$p(\boldsymbol{y}_t|k_t) = \mathcal{N}\left(\boldsymbol{y}_t; g(\boldsymbol{\mu}^{\boldsymbol{x}}_{k_t}, \boldsymbol{\mu}^{\boldsymbol{n}}), \frac{\boldsymbol{\sigma}_{k_t}}{(\mathbf{1} + \exp(\boldsymbol{\mu}^{\boldsymbol{n}} - \boldsymbol{\mu}^{\boldsymbol{x}}_{k_t}))^2} + \frac{\boldsymbol{\sigma}^{\boldsymbol{n}}}{(\exp(\boldsymbol{\mu}^{\boldsymbol{x}}_{k_t} - \boldsymbol{\mu}^{\boldsymbol{n}}) + \mathbf{1})^2}\right) \quad (2.12)$$

が得られる．ただし $\boldsymbol{\sigma}_{k_t}$ はクリーン音声の音響モデルのインデックス k_t に対応する正規分布の分散ベクトルである．

以上では特徴量として FBANK 場合の VTS 近似の計算式を示したが，他にも特徴量として MFCC などを用いたり，g の形を変形して部屋のインパルス応答に関する情報を考慮するようにしても，計算が少し複雑になるだけで，VTS 適応の枠組みそのものはそのまま利用できる [12].

VTS 適応の計算時間の最大の非効率は，音響モデルに含まれる正規分布の個数回，(2.12) の分散の逆行列を求めなければいけないことにある．FBANK を使うと，分散共分散行列が対角になるため大きな問題にはならないが，MFCC を使うと，この分散共分散行列が全角になり，計算量が非常に大きくなってしまう．そこで，あらかじめ音響モデルをクラスタリングしておき，それぞれのクラスタごとに VTS の中心を共有すれば，計算量を削減する手法が提案されている．これをクラス VTS 適応と呼ぶ．クラス VTS 適応を用いると，計算量に余裕ができるため，計算コストのかかる 2 次の VTS や，動的特徴量の適切取り扱いを導入することができるようになり，単純な 1 次の VTS 適応より精度を向上させるような手法も提案されている [13].

2.2.5　特徴量強調（雑音抑圧・音声強調）

雑音環境のミスマッチの低減には，特徴量を音響モデルに近づけようとするアプローチもある．この手法は，ノイジーな音声特徴量をクリーンな音声特徴量に近づけるので，雑音抑圧，音声強調などという用語で呼ばれる．雑音抑圧／音声強調は，音声認識の精度の改善の他にも，ノイジーな音声をクリーンな音声にすることで，ヒトが聞き取りやすいようにする処理にも利用される．

i)　複数マイクロフォンを利用した手法

もし入力として複数のマイクロフォンで収録した音声が得られれば，マイクの位置の違いから得られる，音源の空間情報を利用して，音声と雑音を分離することが可能になる．このような手法としては，適応ビームフォーマ，DUET（Degenerate Unmixing Estimation Technique），ICA（Independent Component Analysis）などがある．

本論文では，複数マイクを用いた手法については詳しく解説せず，以降では，一つのマイクから得られた音声信号を雑音抑圧する手法について見ることにする．

ii)　SS

最も単純なシングルマイクにおける雑音抑圧として，無音区間から推定した雑音のパワースペクトルを，観測したノイジーな音声のパワースペクトルから引き算する，スペクトルサブトラクション（Spectral Subtraction; SS）がある．SS では，観測したノイジー音声のパワースペクトルより，推定した雑音のパワースペクトルの方が大きくなってしまった場合に破綻してしまうので，適切に場合分けを行うことが行われる．

SS では，$\boldsymbol{x}_t$ や $\boldsymbol{y}_t$ の特徴量空間に関する情報を利用していないため，「音声とはおよそこういったものである」という知識や仮定が反映されない．以降，さまざまな方法を使って，この問題点を改良した雑音抑圧手法を紹介していく．

iii)　AFE

さまざまなヒューリスティクスを導入して高精度な特徴量強調を実現する方法として，欧州の標準化団体 ETSI が標準化している AFE（Advanced Front-End）を紹介する（`http://www.etsi.org/WebSite/Technologies/DistributedSpeechRecognition.aspx`）．AFE では，ウィーナーフィルタを二回かけることを中心とした，数々の処理を組み合わせることで特徴量強調を実現している．AFE 処理の中身では，定数が数多く決め打ちされており，「職人技」を導入した雑音抑圧手法であるといえる．AFE は，ETSI のウェブサイトで実装も公開されているので，特徴量強調手法のベースラインとして利用されることが多い [14].

iv)　NMF

対数をとる前のパワースペクトルドメインでは，観測されるパワースペクトルは，音声と雑音が単純な足し算になっている．このパワースペクトルの時系列に対して，NMF（Nonnegative Matrix Factorization）を適用することで，音声と雑音を分離することが可能になる．

Non-negative Matrix Factorization (NMF) とは，$n \times m$ 行列 V が与えられたとき，$V \approx WH$ と分解できるような $n \times r$ の非負行列 W と $r \times m$ の非負行列 H を見つける技術である．今回の場合，パワースペクトルの時系列を V とおく．そのため，n が周波数 bin の数，m が時間である．

NMF では，W も H も同時に推定することができるが，音声強調応用では，W は予め用意しておく．具体的にはまず，クリーン音声データから集めたパワースペクトルを複数並べ，W_{clean}，を用意する．次に，雑音データから集めたパワースペクトルを沢山用意し，W_{noise} を用意する．そして，$W = [W_{\text{clean}} W_{\text{noise}}]$ とおき，NMF のアルゴリズムで，$H = \begin{bmatrix} H_{\text{clean}} \\ H_{\text{noise}} \end{bmatrix}$ を推定する．こうすることで，$W_{\text{clean}} H_{\text{clean}}$ がクリーン音声パワースペクトルの推定値として利用できる [15, 16]．この手法は，exemplar-based method という名称でも知られている．Exemplar-based method では，巨大な W を用いるほど，高い音声認識精度が得られることが知られている．しかしながら，W を大きくすればするほど計算時間も膨大になり，実時間程度での処理はほぼ不可能になってしまう問題がある．

v)　VTS 強調

これまで，パワースペクトルドメインにおける雑音抑圧手法を見てきたが，音声認識応用を考えると，音声波形を得る必要はないので，音声認識の特徴量（パワースペクトラムの log をとった後の特徴量，すなわち FBANK や MFCC など）の段階で雑音抑圧を行えれればよい．本論文では，特徴量の段階での雑音抑圧／音声強調のことを,「特徴量強調」と呼ぶ．

特徴量強調の目標は，ノイジーな音声の特徴量 $\boldsymbol{y}_t$ から，クリーンな音声の特徴量 $\boldsymbol{x}_t$ の推定値 $\hat{\boldsymbol{x}}_t$ を求めることである．以降，特徴量強調手法について詳しく見ていく．

まず，音響モデルの雑音適応でも用いた VTS（Vector Taylor Series）を，特徴量強調に利用する手法を紹介する．VTS 強調は，音響モデルの HMM/GMM とは別に，クリーン音声のモデルを学習し利用することから，Model-Based Feature Enhancement (MBFE) とも呼ばれる．特徴量強調用のクリーン音声のモデルは，音声認識用の音響モデルと比べて小さくできるので，その分，VTS 適応より計算量を削減することができる長所がある．

まず，クリーン音声のモデルとして，$\boldsymbol{x}_t$ が時間によらず一定の K 混合 GMM から出力

されると仮定し，以下のような GMM を学習する．

$$p(k) = \pi_k \tag{2.13}$$
$$p(\boldsymbol{x}_t|k) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k) \tag{2.14}$$

$\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\}_{k=1\cdots K}$ は GMM のモデルパラメタであり，学習データを使って推定する．

特徴量強調を行うときには，まず雑音特徴量が正規分布に従うと仮定し，$p(\boldsymbol{n}) = \mathcal{N}(\boldsymbol{n}, \boldsymbol{\mu^n}, \boldsymbol{\sigma^n})$ のパラメタ $\boldsymbol{\mu^n}, \boldsymbol{\sigma^n}$ を推定する．次に，VTS 適応と同じように VTS 近似を行い (2.12) により $p(\boldsymbol{y}_t|k)$ を求め，

$$p(k|\boldsymbol{y}_t) = \frac{p(k)p(\boldsymbol{y}_t|k)}{\sum_{k=1}^{K} p(k)p(\boldsymbol{y}_t|k)} \tag{2.15}$$

と $p(k|\boldsymbol{y}_t)$ を計算する．

これを用い，VTS 強調では，以下のようにクリーン音声特徴量の推定値を求める [17].

$$\hat{\boldsymbol{x}}_t = \underset{\boldsymbol{x}_t}{\operatorname{argmax}}\, p(\boldsymbol{x}_t|\boldsymbol{y}_t) \tag{2.16}$$
$$= \underset{\boldsymbol{x}_t}{\operatorname{argmax}} \sum_{k=1}^{K} p(\boldsymbol{x}_t, k|\boldsymbol{y}_t) \tag{2.17}$$
$$= \underset{\boldsymbol{x}_t}{\operatorname{argmax}} \sum_{k=1}^{K} p(k|\boldsymbol{y}_t)p(\boldsymbol{x}_t|\boldsymbol{y}_t, k) \tag{2.18}$$
$$\approx \sum_{k=1}^{K} p(k|\boldsymbol{y}_t)g(\boldsymbol{\mu}_k^{\boldsymbol{x}}, \boldsymbol{\mu^n}) \tag{2.19}$$

ただし，(2.19) の近似には別の方法もあり，

$$\hat{\boldsymbol{x}}_t \approx \sum_{k=1}^{K} p(k|\boldsymbol{y}_t) \left(\boldsymbol{\mu}_k^{\boldsymbol{x}} + \frac{\boldsymbol{\sigma}_k^{\boldsymbol{x}}(\boldsymbol{\sigma}_{k_t}^{\boldsymbol{y}})^{-1}}{\mathbf{1} + \exp(\boldsymbol{\mu^n} - \boldsymbol{\mu}_{k_t}^{\boldsymbol{x}})} (\boldsymbol{y}_t - \boldsymbol{\mu}_{\boldsymbol{k_t}}^{\boldsymbol{y}}) \right) \tag{2.20}$$
$$\boldsymbol{\mu}_{k_t}^{\boldsymbol{y}} = g(\boldsymbol{\mu}_{k_t}^{\boldsymbol{x}}, \boldsymbol{\mu^n}) \tag{2.21}$$
$$\boldsymbol{\sigma}_{k_t}^{\boldsymbol{y}} = \frac{\boldsymbol{\sigma}_{k_t}}{(\mathbf{1} + \exp(\boldsymbol{\mu^n} - \boldsymbol{\mu}_{k_t}^{\boldsymbol{x}}))^2} + \frac{\boldsymbol{\sigma^n}}{(\exp(\boldsymbol{\mu}_{k_t}^{\boldsymbol{x}} - \boldsymbol{\mu^n}) + \mathbf{1})^2} \tag{2.22}$$

のような推定値を利用することもできる [18].

vi)　ALGONQUIN

VTS 近似では，クリーン音声モデルのインデックス k_t に対応するクリーン音声の平均 $\boldsymbol{\mu}_{k_t}^{\boldsymbol{x}}$ と，正規分布による雑音モデルの平均値 $\boldsymbol{\mu^n}$ を，VTS 展開の中心としていた．しかしながら，展開の中心が実際の値と離れていると，近似の精度が低下してしまう．

そこで，VTS を行い，それによって推定できたクリーン音声や雑音の推定値を，VTS 近

似の中心としてもう一度 VTS 近似を行う，という繰り返しアルゴリズムを利用することで，さらに精度を向上させられると考えられる．このような近似手法は,「ラプラス近似」のバリエーションとして，機械学習の分野で広く研究が行われている．

ALGONQUIN は，VTS の二次近似を利用し，上記のような繰り返しを行うことで VTS 強調の精度を高めた特徴量強調法である [19]．繰り返しアルゴリズムを使うことによる計算量の増加を防ぐため，特徴量としては FBANK 領域を用いるのが一般的である．定式化の上では，雑音モデルが GMM であってもよいが，計算量の観点から多くの場合正規分布が利用される．

vii)　DNA

Dynamic Noise Adaptation（DNA）では，雑音をガウシアンプロセスでモデル化する [20]．音声に関しては，VTS 強調や ALGONQUIN と同様，FBANK 領域の GMM でモデル化する．また，ALGONQUIN と同様の VTS 近似の繰り返しアルゴリズムを採用している．

VTS 強調，ALGONQUIN，DNA これらすべては，音声特徴量 $\boldsymbol{x}_t$ とノイジーな音声特徴量 $\boldsymbol{y}_t$，雑音特徴量 $\boldsymbol{n}_t$ の生成モデルになっており，さらに VTS 近似によりすべての変数の関係が線形になっている．そのため原理的には，観測したノイジー音声から，雑音特徴量を推定することが可能である．DNA に関する論文では，明示的にグラフィカルモデルが与えられており，雑音特徴量を推定するアルゴリズムが定式化されている．すなわち，予め何らかのアルゴリズムを使って雑音を推定してから特徴量強調を行うのではなく，統一的な枠組みで雑音を推定し，それを特徴量強調に利用することができる．

さらに DNA を用いた特徴量強調の枠組みでは，観測した音声がクリーン音声に近い場合には特徴量強調を行わず，雑音が多く重畳している場合のみに特徴量を強調を行うように，選択をする手法が提案されており，DNA-CD（Condition Detection）と呼ばれている [21]．DNA-CD により，入力がクリーン音声だった場合の性能の劣化がなくなり，商品として売られているグレードの音声認識システムの性能を，どのような雑音環境下でも向上させられることが示されている．

viii)　SPLICE

次に，VTS 系のアルゴリズムから離れ，$\boldsymbol{y}_t$ と $\boldsymbol{x}_t$ の関係を，時間的対応のとれている学習データ（ステレオデータ）を学習データとして学習して利用する方法を紹介する．SPLICE（Stereo Piecewise Linear Compensation for Environments）は，ステレオデータを用いる特徴量強調法としてもっとも有名なものである [22]．このような方法としては，他にも SSM（Stereo-based Stocastic Mapping）[23] や MEMLIN（Multi-Environment Model-based Linear Normalization）[24] などがある．

ここではこのステレオデータを $\boldsymbol{X} = \{\boldsymbol{x}_t\}_{t=1\cdots T}$，$\boldsymbol{Y} = \{\boldsymbol{y}_t\}_{t=1\cdots T}$ とおく．まず，$\boldsymbol{y}_t$ が時間によらず一定の K 混合 GMM から出力されると仮定し，

$$p(k) = \pi_k \tag{2.23}$$

$$p(\boldsymbol{y}_t|k) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k) \tag{2.24}$$

とおく．ここで，GMM のパラメタ $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\}_{k=1\cdots K}$ は，$\boldsymbol{Y}$ から学習する．

SPLICE では，$\boldsymbol{y}_t$ の GMM パラメタを利用し，以下のように雑音抑圧結果 $\hat{\boldsymbol{x}}_t$ を得る．

$$\hat{\boldsymbol{x}}_t = \sum_{k=1}^{K} p(k|\boldsymbol{y}_t)\boldsymbol{A}_k \begin{bmatrix} 1 \\ \boldsymbol{y}_t \end{bmatrix} \tag{2.25}$$

ここで，$p(k|\boldsymbol{y}_t)$ は

$$p(k|\boldsymbol{y}_t) = \frac{p(k)p(\boldsymbol{y}_t|k)}{\sum_{k=1}^{K} p(k)p(\boldsymbol{y}_t|k)} \tag{2.26}$$

$$= \frac{\pi_k \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)} \tag{2.27}$$

と計算できる．$\boldsymbol{A}_k$ は，ステレオデータから学習する．学習法にはさまざまな方法があるが，一番簡単なのは重み付き最小二乗誤差を最小化するように決定することで，以下で定式化される．

$$\underset{\boldsymbol{A}_k}{\operatorname{argmin}} \sum_t p(k|\boldsymbol{y}_t) \left|\left| \boldsymbol{x}_t - \boldsymbol{A}_k \begin{bmatrix} 1 \\ \boldsymbol{y}_t \end{bmatrix} \right|\right|^2 \tag{2.28}$$

これの解析解は，

$$\boldsymbol{A}_k = \overline{\boldsymbol{X}}\boldsymbol{P}_k\overline{\boldsymbol{Y}}^\top \left(\overline{\boldsymbol{Y}}\boldsymbol{P}_k\overline{\boldsymbol{Y}}^\top\right)^{-1} \tag{2.29}$$

となる．ただし，$\overline{\boldsymbol{X}}$，$\overline{\boldsymbol{Y}}$ はそれぞれ $\boldsymbol{x}_t$，$\begin{bmatrix} 1 \\ \boldsymbol{y}_t \end{bmatrix}$ を学習データのフレーム数分並べた行列，$\boldsymbol{P}_k$ は $p(k|\boldsymbol{y}_t)$ を学習データのフレーム数分並べた物を対角成分にもつ対角行列とする．

SPLICE による特徴量強調を簡略化したものを 図2.1 に示す．そして、それぞれの部分空間の中では，線形変換 $\boldsymbol{A}_k$ は一定である．部分空間のインデックス k が変われば，変換はまったく別のものが利用される．これにより，結果的に，全体としては非線形変換を表現している．すなわち SPLICE は，区分的線形変換（piecewise linear transformation）を使ってノイジーな音声の特徴量をクリーンな音声の特徴量の推定値に変換している．

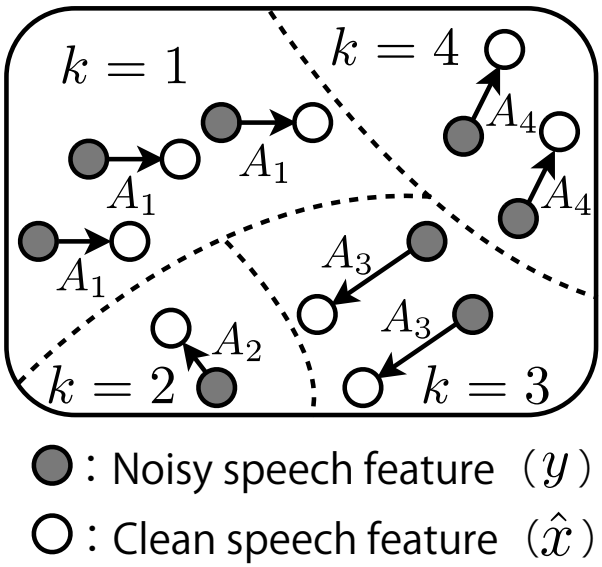

図 2.1: SPLICE によるノイジーな特徴量 $\boldsymbol{y}$ からクリーン音声特徴量の推定値 $\hat{\boldsymbol{x}}$ への変換の概念図．区分的線形変換により，全体では非線形な変換を実現している．k は部分空間のインデックスで，図では部分空間の数は 4 としている．分かりやすさのため，$p(k|\boldsymbol{x}_t)$ が，ある k で 1 となり，それ以外で 0 となるような近似を行い，さらに線形変換は足し算演算（二次元空間上でのシフト）のみとした．実際は区分化はソフトに行われ，線形変換回転など含む任意の変換が行われる．

なお，SPLICE に関しては，著者が作成した matlab コードが公開されている[2]．

ix)　NMN-SPLICE

SPLICE の学習には，クリーン音声特徴量とノイジー音声特徴量のステレオデータが必要になるが，もし，このノイジー音声特徴量の雑音環境と，実際に音声認識を行うときの雑音環境が異なると，SPLICE ではミスマッチを低減しきれず，精度が低下してしまう問題が起こる．

この問題はステレオデータを用いる特徴量強調手法の本質的な問題であるが，これをある程度低減する方法として，Noise Mean Normalization (NMN) を導入した SPLICE が提案されている [22]．これは，SPLICE の処理を行う前に，すべての特徴量から雑音特徴量の推定値 $\hat{\boldsymbol{n}}$ を引き算し，SPLICE を行った後　$\hat{\boldsymbol{n}}$ を足し算する処理である．

つまり，$\boldsymbol{y}_t$ の GMM ではなく，$\boldsymbol{y}_t - \hat{\boldsymbol{n}}$ の GMM を学習しておき，

$$\hat{\boldsymbol{x}}_t - \hat{\boldsymbol{n}} = \sum_{k=1}^{K} p(k|\boldsymbol{y}_t - \hat{\boldsymbol{n}}) \boldsymbol{A}_k \begin{bmatrix} 1 \\ \boldsymbol{y}_t - \hat{\boldsymbol{n}} \end{bmatrix} \tag{2.30}$$

[2] https://sites.google.com/site/suzukimasayuki/splice

といった式で特徴量強調を行う．

NMN-SPLICE を用いると，特に無音区間において，$\boldsymbol{y}_t - \hat{\boldsymbol{n}} = 0$ が近似的に成り立つことから，$\boldsymbol{y}_t$ より，$\boldsymbol{y}_t - \hat{\boldsymbol{n}}$ の方が，特徴量空間における広がりは狭いと考えられる．そのため，NMN-SPLICE の方が SPLICE と比べて性能が高くなると考えられ，実際実験的に効果が確かめられている．

2.2.6　Uncertainty decoding

雑音環境のモデル適応と特徴量強調を比較すると，一般的には，精度の面ではモデル適応が，計算速度の面では音声強調が良い，という関係がある．音響モデル適応では，音素などの音響モデルの単位ごとに異なる処理を行うために，より適切な処理が実現できるがその分計算時間が増加してしまう．一方特徴量強調では音素などの情報は強調時には使わないために，高速な処理が実現できる．

Uncertainty decoding（UD）では，モデル適応に近い精度を，特徴量強調と同程度の計算量で得ようとするための手法である．考え方としては，特徴量強調結果がうまく入ったか，いかなかったかの uncertainty を，後段の音響モデルをつかったデコーディング処理まで伝えよう，というものである．

UD では，雑音環境の音響モデル適応で使った $p(\boldsymbol{y}_t|\boldsymbol{x}_t, k)$ が，k に非依存であると仮定する．すなわち，

$$
\begin{aligned}
p(\boldsymbol{y}_t|k) &= \int p(\boldsymbol{y}_t, \boldsymbol{x}_t|k) d\boldsymbol{x}_t && (2.31)\\
&= \int p(\boldsymbol{y}_t|\boldsymbol{x}_t, k) p(\boldsymbol{x}|k) d\boldsymbol{x}_t && (2.32)\\
&\approx \int p(\boldsymbol{y}_t|\boldsymbol{x}_t) p(\boldsymbol{x}|k) d\boldsymbol{x}_t && (2.33)\\
&= \int \frac{p(\boldsymbol{x}_t|\boldsymbol{y}_t) p(\boldsymbol{y}_t)}{p(\boldsymbol{x}_t)} p(\boldsymbol{x}_t|k) d\boldsymbol{x}_t && (2.34)
\end{aligned}
$$

のように $p(\boldsymbol{y}_t|k)$ が計算できると仮定する．この近似により，雑音のモデル適応で問題となっていた，各音響モデルごとの計算がなくなり，計算量は特徴強調と同程度まで下がる．

ここで，$p(\boldsymbol{x}_t|\boldsymbol{y}_t), p(\boldsymbol{x}_t), p(\boldsymbol{y}_t)$ を計算する必要がある．これらを求めるために，UD と組み合わせる特徴量強調法としては，ステレオデータに基づく SPLICE などの手法が利用される．特に，$\boldsymbol{x}_t$ と $\boldsymbol{y}_t$ の結合ベクトルを GMM でモデル化する特徴量強調法を UD と組み合わせると，SPLICE with UD より高い精度が得られることが知られており，これは Joint UD（JUD）と呼ばれている [25]．

2.2.7　特徴量正規化

これまでは，話者の違いにしろ雑音環境の違いにしろ，音響モデルを特徴量に近づけたり，特徴量を音響モデルに近づけたりといった処理を紹介してきたが，それとは少し異なるアプローチとして，特徴量を正規化することでミスマッチを低減させるアプローチを紹介する．

i)　CMN

特徴量正規化で最も有名なのは，特徴量の時間平均を 0 に正規化する方法で，Cepstrum Mean Normalization (CMN) と呼ばれている．CMN はシンプルかつ高い効果があるため，デファクトスタンダードとして広く利用されている．また，音声認識を逐次的に行うために，短い時間範囲で時間平均を 0 にするオンライン版の CMN も広く利用されている．

ii)　CMVN

CMVN とは Cepsturm Mean and Variance Normalization の略で，文字通り，平均を 0 に揃えるだけでなく，分散を 1 に正規化する手法である．データ量が沢山ある場合には，CMN より高い性能を示すことがある．

iii)　HEQ

Histogram EQuarization（HEQ）では，特徴量のヒストグラムを何らかの分布に正規化する手法である．具体的には，正規分布を使うことが多い．HEQ は簡単ながらも，特に雑音環境下では非常に高い性能を示す [26]．

2.2.8　音響モデルの話者・雑音適応学習

特徴量正規化は，音響モデルの学習データにも，評価データにも，同じ正規化処理を施すことで最終的なミスマッチを減らす方法である．ここで，特徴量ドメインの話者適応や特徴量強調も，正規化の一種とみなせば，評価データだけでなく，音響モデルの学習データにも同じ処理をかけて音響モデルを学習してやることで，よりミスマッチを低減できる．この処理は，話者適応の場合には Speaker Adaptive Training（SAT）[27]，雑音適応の場合には Noise Adaptive Training（NAT）と呼ばれている [28]．

2.2.9　音響モデルの識別学習

音響モデルである HMM/GMM のパラメタ $\mathbf{\Lambda}$ とおく．$\mathbf{\Lambda}$ には，GMM の各コンポーネントの正規分布の平均ベクトル，分散ベクトル，GMM のコンポーネント重み，各 HMM

の状態遷移確率，などが含まれている．テキストラベル付きの音声データを使えば，$\mathbf{\Lambda}$ をバウムウェルチアルゴリズムを使って最尤（Maximum Likelihood ML）推定することができる．すなわち，

$$\mathbf{\Lambda}_{\mathrm{ML}} = \operatorname*{argmax}_{\mathbf{\Lambda}} \log p(\boldsymbol{X}|W, \mathbf{\Lambda}) \tag{2.35}$$

を解くことでパラメタを推定を行なっている．

しかしここで，学習データが出力される尤度がもっとも大きくなるようにパラメタを学習することと，音声認識の精度を高くすることは，必ずしも一致しない．そこで，HMM/GMM のパラメタ Λ を，音声認識の精度が高くなるような基準で学習しようとする手法が数多く提案されている．これは，音響モデルの識別学習と呼ばれている．音響モデルの識別学習に関する最新のサーベイには，[29] がある．

i)　MMI 基準

音声認識の目的は，音声特徴量系列 $\boldsymbol{X}$ が与えられたときに，話された内容の単語系列 W を推定することなので，$\boldsymbol{X}$ が与えられた時の W の相互情報量が最大化されるように $\mathbf{\Lambda}$ を学習することを考える．この基準は Maximum Mutual Information (MMI) 基準と呼ばれる．

$$\begin{aligned}
\mathbf{\Lambda}_{\mathrm{MMI}} &= \operatorname*{argmax}_{\mathbf{\Lambda}} \log p(W|\boldsymbol{X}, \mathbf{\Lambda}) && (2.36)\\
&= \operatorname*{argmax}_{\mathbf{\Lambda}} \log p(W, \boldsymbol{X}|\mathbf{\Lambda}) - \log p(\boldsymbol{X}|\mathbf{\Lambda}) && (2.37)\\
&= \operatorname*{argmax}_{\mathbf{\Lambda}} \log p(\boldsymbol{X}|W, \mathbf{\Lambda}) + \log p(W|\mathbf{\Lambda}) - \log \sum_{W'} p(W', \boldsymbol{X}|\mathbf{\Lambda}) && (2.38)\\
&= \operatorname*{argmax}_{\mathbf{\Lambda}} \log p(\boldsymbol{X}|W, \mathbf{\Lambda}) - \log \sum_{W'} p(\boldsymbol{X}|W', \mathbf{\Lambda}) p(W') && (2.39)
\end{aligned}$$

ただし (2.38) から (2.39) では，W の出現確率は音響モデルのパラメタ $\mathbf{\Lambda}$ とは独立であると仮定し，$\mathbf{\Lambda}$ に関する最大化では無視できることを利用した．

(2.39) の第二項は，全 W' について考慮するのは計算コスト的に難しいので，いったん ML 基準などで学習した HMM/GMM を用いて音声認識を行なって，その結果得た上位 N 個の仮説や，もしくは認識結果のラティスに範囲をしぼって計算する．また，この仮説やラティスを得るときの言語モデルとしては，1-gram のような弱い言語モデルを用いた方が良いことが経験的に知られている．そしてこの最大化問題は，(2.39) を $\mathbf{\Lambda}$ のそれぞれのパラメタで偏微分及び二回偏微分した値を用いた逐次更新アルゴリズムで最適化する．

なお，デコーディングの際に音響モデルのスコアと言語モデルの対数スコアに定数をかけて調整したように，識別学習においても $\log p(\boldsymbol{X}|W', \mathbf{\Lambda})$ を定数倍することも一般的に行われているが，ここでは省略している．

ii)　MPE 基準

(2.39) を見ると，左辺は $\mathbf{\Lambda}$ と正解の単語系列が与えられた下での $\boldsymbol{X}$ の対数尤度で，右辺は W' が正解しかとりうらない場合には，左辺と同じになり，全体では打ち消し合って 0 になる（すなわち $\mathbf{\Lambda}$ が更新されない）．一方 W' が間違った単語系列になっている場合は，左辺との差が大きくなり，全体は大きくなる（すなわち $\mathbf{\Lambda}$ を大幅に更新する）．

MMI 基準では,「さとう」を,「さいとう」と間違えようと,「すずき」と間違えようと，どちらも同じペナルティーを与えることになる．ここで，音響モデルでは単語系列 W より細かい，音素や状態をモデル化している．そのため,「さとう」を「さいとう」と間違えるたった 1 音素 /i/ の挿入誤りよりも,「さとう」を「すずき」と間違えるのでは，よりペナルティーが大きいと考えられる．そこで，MMI 基準に変更を加え，音素などのより細かい単位の誤りを最小化する目的関数を利用して識別学習を行うことで，より高い精度を実現できることが知られている [30].

音素誤り率を最小化する Minimum Phone Error (MPE) 基準での音響モデルパラメタ $\mathbf{\Lambda}$ の学習は，以下のように定式化できる．

$$\mathbf{\Lambda}_{\mathrm{MPE}} = \operatorname*{argmax}_{\mathbf{\Lambda}} \sum_{W'} \log p(W', \boldsymbol{X}|\mathbf{\Lambda}) A(W, W') - \log \sum_{W'} p(W', \boldsymbol{X}|\mathbf{\Lambda}) p(W') \quad (2.40)$$

ここで $A(W, W')$ は，W' が W として音素的にはどれだけ正しいかの Accuracy を表す関数である．以降の式展開方や解き方は，音響モデルスコアの定数倍などは，MMI 基準の場合と同様である．

iii)　BMMI 基準

MPE よりさらに精度が高くなる識別学習の基準として，Boosted MMI （BMMI）というものが知られている [31]．BMMI は以下の式で定式化できる．

$$\mathbf{\Lambda}_{\mathrm{BMMI}} = \operatorname*{argmax}_{\mathbf{\Lambda}} \log p(W, \boldsymbol{X}|\mathbf{\Lambda}) - \log \sum_{W'} p(W', \boldsymbol{X}|\mathbf{\Lambda}) p(W') \exp\left(-bA(W, W')\right) \quad (2.41)$$

ここで b は定数であり，バリデーションセットを用いて決定する．この式は，MMI 基準の左辺に $\exp\left(-bA(W, W')\right)$ をかけたものとなっており，この項により，競合する仮説の中でも特に正解精度の高い仮説に，ペナルティを与えることになる．これは，結果的に識別境界からのマージンを大きくするような学習になっている．

この考え方は，MPE にも導入することができ，“BMPE” 基準も用いることができるが，実用的には BMMI が利用されるケースが多いようである．

iv)　I-smoothing

HMM/GMM の識別学習を実際に行うと，簡単に過学習してしまうことが知られている．そこで，パラメタを推定する際に，ML 推定のパラメタを使って smoothing を行うことで，過学習を抑圧する I-smoothing と呼ばれる手法が広く利用されている [30]．

2.2.10　特徴量の識別学習

ここまで，音響モデルである HMM/GMM のパラメタを識別学習する手法をみてきたが，特徴量を，音声認識率が高くなるように変換する手法もある．

i)　LDA

MFCC の Δ 特徴量は，MFCC 数フレーム分の線形変換（中心フレームにおける時間方向の回帰）によって計算されている．ここで，MFCC 数フレーム分の線形変換を，LDA（Linear Discriminant Analysis）を用いた次元圧縮に置き換えることで，特徴量抽出を識別的に行うことができ，精度が向上することが知られている．

LDA の入力としては，例えば前後 4 フレームずつ，合計 9 フレーム程度の MFCC を用いる．教師ラベルには，いったん MFCC + Δ + $\Delta\Delta$ を用いて学習した HMM/GMM を使ってアライメントをとり，その HMM の状態 ID を利用する．LDA による次元圧縮後の特徴量としては，およそ 40 次元程度が利用される．

LDA は，各クラスが正規分布である場合に最適な教師あり次元圧縮法であるが，実際には各クラスの特徴量は正規分布に従っているとは言い切れないため，LDA とは別の教師あり次元圧縮法が用いられることもある．具体的には，Heteroscedastic LDA（HLDA）などが利用される [32, 33]．

ii)　STC

STC（Semi-Tied Covariance）は厳密には特徴量の識別学習を行う手法ではないが，前述の LDA と組み合わせて利用されることが多いのでここで紹介する [34]．

音響モデルの HMM/GMM の各正規分布の分散共分散行列は，対角であることが仮定されている．そこで，上記の LDA によって抽出された特徴量が，対角の HMM/GMM から出力されるという近似精度がより上がるように特徴量空間を回転させるのが，STC と呼ばれる手法である．STC により，音声認識の精度が向上することが知られている．

iii)　fMPE/fBMMI

次に，fMPE や fBMMI など，音響モデルの識別学習で使われた基準に “f” を頭に付ける名称を持つ特徴量の識別学習法を紹介する [35]．

まず，特徴量空間上で，沢山のガウス分布を用意する．具体的には，音響モデルの HMM/GMM の GMM の各正規分布をすべて使っても良いし，SPLICE などのように，GMM を別に学習してもよい．合計 K 個のガウス分布が得られれば，ガウス分布のインデックスを k として，$\{p(\boldsymbol{x}_t|k)\}_{k=1\cdots K}$ が計算できる．また，時刻 t 前後数フレームの特徴量からも，同様の値が計算できる．例えば [35] では，$\boldsymbol{x}_t$ と，$(\boldsymbol{x}_{t+1}+\boldsymbol{x}_{t+2})/2$，$(\boldsymbol{x}_{t+3}+\boldsymbol{x}_{t+4}+\boldsymbol{x}_{t+5})/3$，$(\boldsymbol{x}_{t+6}+\boldsymbol{x}_{t+7}+\boldsymbol{x}_{t+8}+\boldsymbol{x}_{t+9})/4$，およびこれらの逆方向 $(\boldsymbol{x}_{t-1}+\boldsymbol{x}_{t-2})/2\cdots$ を使う．これらの事後確率をすべてまとめて，$\boldsymbol{h}_t$ なるベクトル表現する．$\boldsymbol{h}_t$ の次元数は，[35] の場合 $K\times 7$ となる．そして，以下のような線形変換により，新しい特徴量 $\boldsymbol{y}_t$ を抽出することを考える．

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{M}\boldsymbol{h}_t \tag{2.42}$$

$\boldsymbol{M}$ は，音声認識の精度が高くなるよううな目的関数を利用して最適化される．目的関数としては，HMM/GMM の識別学習にも利用した，MPE や BMMI などが利用できる．どの目的関数を使ったかによって，それぞれ fMPE や fBMMI などと呼ばれる．

fMPE や fBMMI は，HMM/GMM の識別学習と同等の精度向上を示し，その上それらを組み合わせることでさらなる精度向上が実現できることが知られている．そのため特徴量の識別学習は非常によく利用されている．

(2.42) と SPLICE の (2.30) を見比べると，およそ同じような方法で変換が行なわれている．実は，SPLICE では，ステレオデータの最小二乗誤差基準で学習を行なっていたが，SPLICE の学習基準を MPE や BMMI にしてしまえば，両者はまったく同じものである [36].

2.2.11　識別モデルによる音声認識

ここまで HMM/GMM の識別学習と，特徴量の識別学習を用いていたが，結局のところ音響モデルとして HMM/GMM を利用している．ここで HMM/GMM による音響モデルには，同じフレーム内で特徴量が独立であることを仮定しているなどといった，実際の音声では成り立っていないような仮定が使われてしまっている．そこで，HMM/GMM を識別学習するのではなく，音声認識全体を識別モデルを用いて行う手法がいくつか提案されている．

音響モデルとして HMM/GMM を使う場合には，単語系列 W と観測データ $\boldsymbol{X}$ の同時確率をモデル化した，いわゆる「生成モデル」が利用されている．すなわち音声認識は以下で定式化される．

$$\hat{W} = \operatorname*{argmax}_W p(W,\boldsymbol{X}) \tag{2.43}$$

$$= \operatorname*{argmax}_W p(\boldsymbol{X}|W)p(W) \tag{2.44}$$

そして $p(\boldsymbol{X}|W)$ を HMM/GMM でモデル化し，$p(W)$ を N-gram でモデル化していた．

一方識別モデルを使う音声認識は，以下のように定式化される．

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, p(W|\boldsymbol{X}) \tag{2.45}$$

そして，$p(W|\boldsymbol{X})$ を，ログリニアモデルなどでモデル化する．

識別モデルによる音声認識では，生成モデルによる音声認識と異なり，言語的制約を導入するのが難しい．そのため，例えば $p(W|\boldsymbol{X})$ を CRF（Conditional Random Field）でモデル化するアプローチは，音素認識で高い精度が実現できても，大語彙音声認識には利用しにくい問題がある．また，HMM/GMM では利用できた話者・雑音適応なども利用できない問題もある．

そこで実用的には，生成モデルを用いてある程度候補となる仮説空間を絞り，その仮説空間を入力とした識別モデルを使って音声認識を行うアプローチが広く利用されている．これにより，HMM/GMM を利用した音声認識を行う際にこれまで紹介した話者・雑音適応もそのまま利用されるため，頑健性を保ったまま，識別モデルによる精度向上を実現することができる．このようなアプローチとしては，SCARF（Segmental CRF を使った方法）[37] [3]や，Gales らの一連の研究 [38] などがある．

2.2.12　タンデムアプローチ

タンデムアプローチとは，音素などの事後確率を特徴量として用いる手法である [39]．まず，MFCC の前後数フレームを入力として，音素（もしくは HMM の状態）ラベルを出力するモデルを学習する．この学習用のラベルは，単語ラベル音声データを強制アライメントすることで得る．そしてフレーム単位で音素事後確率を求め，それを音声認識の特徴量として利用する．音素事後確率になった特徴量ベクトルは，非常にスパースになっており，これを使って音声認識を行うと精度が向上することが経験的に知られているため，注目されている．

フレーム単位の音素事後確率を計算するモデルとしては，ニューラルネットワークが広く利用されている．また，実際には音声認識利用しにくい識別モデルとして紹介した CRF なども，タンデムアプローチの要素技術としては実用的に利用可能である．また，次に述べる DNN も，タンデムアプローチのために利用することができる．

[3]SCARF の開発者は，最終的には HMM/GMM のような生成モデルは必要なくなる，と主張している．しかしながら現実的には，HMM/GMM を導入しないと高い精度が得られない．

2.2.13　DNN を用いた音響モデル

近年，HMM/GMM の GMM の代わりに，DNN（Deep Neural Network）を利用することで，非常に高い精度を実現できるとして注目が集まっている [40].

ニューラルネットワークの学習では，適当な初期値を定めたあと，教師データを用いてバックプロパゲーションを用いてネットワークの重みを更新していく．ここで隠れ層が 1 層のニューラルネットワークの場合，初期値がランダムに設定されていても，ある程度よい性能を持つニューラルネットワークが作成できることが経験的に知られている．隠れ層の数を増やしていくと，モデルとしての複雑度が上がり，より複雑な現象をモデル化できると考えられる．しかし，隠れ層を複数にした場合，初期値をランダムな値にしてバックプロパゲーションで学習しようとすると，過学習の問題が発生し，性能が低下してしまうことが知られていた．

Hinton らはこの問題に対し，Restricted Boltzmann Machine (RBM) を何層も重ねた Deep Belief Network (DBN) を初期値として利用する解決法を提案した [41]．DBN を初期値としてバックプロパゲーションを行うことで，最終的に得られる DNN が，従来の state-of-the-art の性能を大きくこえる精度を実現できることが示され，非常に高い注目を集めている．

RBM は，観測データ $\boldsymbol{x}$ と，隠れ層 $\boldsymbol{h}_l$ を一つもつような生成モデルである．$\boldsymbol{x}$ と $\boldsymbol{h}_l$ は 0 か 1 のみの値を持つベクトルである[4]．RBM には，入力層と隠れ層の値を変数とする以下のようなエネルギーが定義されている．

$$E(\boldsymbol{x}, \boldsymbol{h}) = -\sum_i b_i^x x_i - \sum_j b_j^h h_j - \sum_{i,j} w_{ij} x_i h_i \tag{2.46}$$

$\{b_i^x, b_j^h, w_{ij}\}_{i,j}$ が RBM のパラメタであり，それぞれ入力層のバイアス，隠れ層のバイアス，ネットワークの重みの i, j, ij に対応する要素である．x_i, h_j は，それぞれ $\boldsymbol{x}$ の i 番目の要素，$\boldsymbol{h}$ の j 番目の要素を表す．入力 $\boldsymbol{x}$ が与えられた時，h_j が 1 になる確率は $\sigma(b_j^h + \sum_i x_i w_{ij})$（$\sigma$ はシグモイド関数）で与えられる．今度は逆に，$\boldsymbol{h}$ が与えられた時に x_i が 1 になる確率も出すことができて，それは $\sigma(b_i^x + \sum_j h_j w_{ij})$ となる．

RBM は，複数の観測データが与えられたときに，ネットワーク重みを教師なしで学習することができる．パラメタの更新式は，データが与えられたとき，$\boldsymbol{x}$ から $\boldsymbol{h}$ の各要素が 1 になる確率を出したときに x_i と h_j が同時に 1 となる割合と，$\boldsymbol{x}$ から $\boldsymbol{h}$ をサンプリングし，その $\boldsymbol{h}$ から $\boldsymbol{x}$ の各要素が 1 になる確率を出したときに x_i と h_j が同時に 1 となる割合が，同じになるように決められる．

以上の方法で一層分の RBM が学習できたら，$\boldsymbol{x}$ と $\sigma(b_j^h + \sum_i x_i w_{ij})$ から得られる $\boldsymbol{h}$ を入力と思って，もう一層分 RBM を学習する．これを繰り返して，任意層の DBN を構築

[4] $\boldsymbol{x}$ は 0 もしくは 1 しかとらないとおいたが，入力や出力が連続値をとる場合には，あらかじめデータを分散 1，平均 0 となるように規格化しておけば，同様の手法が利用できる．

する．最後に，教師データと DBN を用い，バックプロパゲーションでパラメタを更新すれば，性能のよい DNN を得ることが出来る．

RBM の実装の細かい設定に関しては，[42] が詳しい．さらに近年，dropout [43] と呼ばれる技術が発明され，後にのべるシステムコンビネーションの効果を簡易な処理で実現できることが示され，さらなる精度向上が実現できることが示されている．近い将来，音声認識の音響モデルとしても，dropout が利用されると考えられる．

音声認識の音響モデルとして HMM/DNN を用いる場合，音素や状態のインデックスを k，その数を K として，DNN は入力層が特徴量，出力層が K 個あり，それぞれが音素や状態インデックスの確率を表すように学習する．すなわち，$p(k|\boldsymbol{x}_t)$ を学習する．HMM と一緒に用いる場合には，$p(\boldsymbol{x}_t|k)$ を計算する必要があるので，これは以下のように計算する．

$$p(\boldsymbol{x}_t|k) = \frac{p(k|\boldsymbol{x}_t)p(\boldsymbol{x}_t)}{p(k)} \tag{2.47}$$

ここで，$p(k|\boldsymbol{x}_t)$ は DNN から計算し，$p(k)$ は単純に学習データをカウントすることで学習する．$p(\boldsymbol{x}_t)$ は，音声認識を行う際には定数項になるため，無視できる．

また，バックプロパゲーションを用いて音響モデルとしての HMM/DNN のパラメタを学習する際には，フレーム単位ではなく，文全体の音声認識の誤りを減らすような基準を用いた方が精度が高くなることが知られている [44].

2.2.14　クラス言語モデル

これまではすべて音響モデルに関する要素技術を紹介してきたが，次からは言語モデルに関する要素技術を紹介する．

言語モデルとしては，単語の N-gram が広く用いられているが，単語ではなく，もう少し広い概念である単語のクラスの N-gram を使う手法がある．クラス N-gram を使った手法で，単語 N-gram 精度を越えたモデルとして，Model M が広く知られている [45]．$N = 3$ の場合の Model M は，以下のようなモデルである．

$$p(w_1, w_2, \cdots, w_I) = \prod_{i=1}^{I+1} p(c_i|w_{i-2}, w_{i-1}) \prod_{i=1}^{I} p(w_i|w_{i-2}, w_{i-1}, c_i) \tag{2.48}$$

ただし c_i は w_i の単語クラスで，$c_i(w_i)$ は予め単語クラスタリングを行い，一意に定まるようにしておく．(2.48) の後半の項が，注目する単語のクラス c_i を既知とした場合の，通常の単語 N-gram 確率と同じものになる．加えて前半の項は，単語の履歴から，今のクラスが出現する確率である．

2.2.15　ニューラルネットワークを用いた言語モデル

N-gram を置き換えようとする言語モデルの研究の中では，ニューラルネットワーク（Neural Network）を使った言語モデル（Language Model）である NNLM が，精度が高いことで知られている [46]．また，NNLM を “deep” にすることで精度を高めた研究もある [47]．

NNLM の問題点としては，ある単語の出現確率が，文頭からすべての単語がないと求められないことがある．一方 N-gram や Model M では，その直前 $N-1$ 個の単語のみさえあれば，出現確率が求められる．そのため，NNLM を言語モデルとして用いる場合，デコーディングを行う際に，膨大な計算量が必要になってしまう．そのため NNLM による単語系列の出現確率は N-gram や NNLM を置き換えるものとはなっておらず，N-gram の補完に用いられたり，次に述べる識別的言語モデルの特徴量として使うのが実用的である．

2.2.16　識別的言語モデル

いったん HMM/GMM もしくは HMM/DNN，単語 N-gram もしくは Model M などを用いて音声認識を行えば，上位 N 個の仮説，もしくは単語ラティスやコンフュージョンネットワークを出力することができる．識別的言語モデルは，この N-best リストや，単語ラティスやコンフュージョンネットワークを入力として，言語的な特徴量を用いてそのリランキングを行うための識別モデルである [48]．

識別的言語モデルの特徴量としては，例えば先の NNLM から計算した $p(W)$ や，計算量の都合で利用できなかった N の大きい N-gram などが利用できる．他にも，Bag of Words（BOW）などといった，任意の特徴量が利用できる．識別的言語モデルの学習データは例えば N-best リストとして与えられるため，単純な正解・不正解を決めるのではなく，ランキングを行うことが必要になる．そこで，識別モデルのパラメタの学習基準として，順位が上になるものを常に正解として扱う Round-Robin Duel Discrimination（R2D2）[49] などのような，さまざまな基準が提案されている．また，学習データは，いったん音声認識をすることで生成することができるが，その他にも，擬似的な音声認識を行うことでデータを生成する研究も行われている [50]．

識別的言語モデルと，先に紹介した識別モデルによる音声認識は，本質的には同じものである．あえて違いを述べれば，識別的言語モデルは言語情報を積極的に利用しようとし，時間的に広い範囲にまたがる特徴量が利用可能なモデルを利用する傾向にあるのに対し，識別モデルによる音声認識では，音響情報を積極的に利用とし，フレーム，音素，単語といった細かい単位で処理を行なう傾向にある，という違い程度である．

2.2.17　システムコンビネーション

複数のシステムから得られた複数の音声認識結果を組み合わせることで，単独のシステムの精度より高い精度を得ることができる．音声認識のシステムコンビネーションとしては，投票によって結果を決める ROVER（Recognize output Voting Error Reduction）やコンフュージョンネットワークの上でコンビネーションを行う CNC（Confusion Network Combination）などがある [51].

2.2.18　最新の話題

ここまで，音声認識の精度を高めるための現時点で最も新しい技術を紹介してきたが，音声認識技術は今後も発展を続けていくと考えられるため，常にサーベイを続ける必要がある．音声認識の有名な国際会議に ICASSP と INTERSPEECH がある．また，音声認識の会議に限らず，ICML や NIPS などの機械学習のカンファレンスでも，音声認識に利用できる技術が発表されることがある．

また，例えば以下のパブリケーションリストなども，最新のトレンドを追うのに有効である．

- Mark Gales（`http://mi.eng.cam.ac.uk/~mjfg/publications.html`）
- Daniel Povey（`https://sites.google.com/site/dpovey/my-publications`）
- Geoffrey Hinton（`http://www.cs.toronto.edu/~hinton/papers.html`）
- Microsoft チーム（`http://research.microsoft.com/en-us/groups/srg/papers.aspx`）
- Google チーム（`http://research.google.com/pubs/SpeechProcessing.html`）

当然，これらの他にも注目すべき研究者や研究チームはたくさん存在しているし，これから新しい研究者も増えているので，ぜひそれ以外の論文も調査されたい．上記のウェブサイトは，調査をはじめる足がかりとして利用されたい．

2.3　既存の音声認識システム

実際に音声認識システムを実装・運用する際には，これまで述べてきた要素技術を適当に組み合わせてシステムを構築することになる．本節では，実際に利用されている，いくつかの音声認識システムを紹介する．

2.3.1　IBM のアラビア語音声認識システム

2009 年に発表された，IBM による，DARPA GALE Program の成果物であるアラビア語の音声認識システムを紹介する [52]．2009 年時点では，HMM/DNN による音響モデル，

Model M による言語モデルが利用されはじめた頃で，このシステムでは利用されていない．しかし，それ以外の要素技術は，2009 年時点でオリジナルの手法は提案されおり，このシステムでも利用されている．

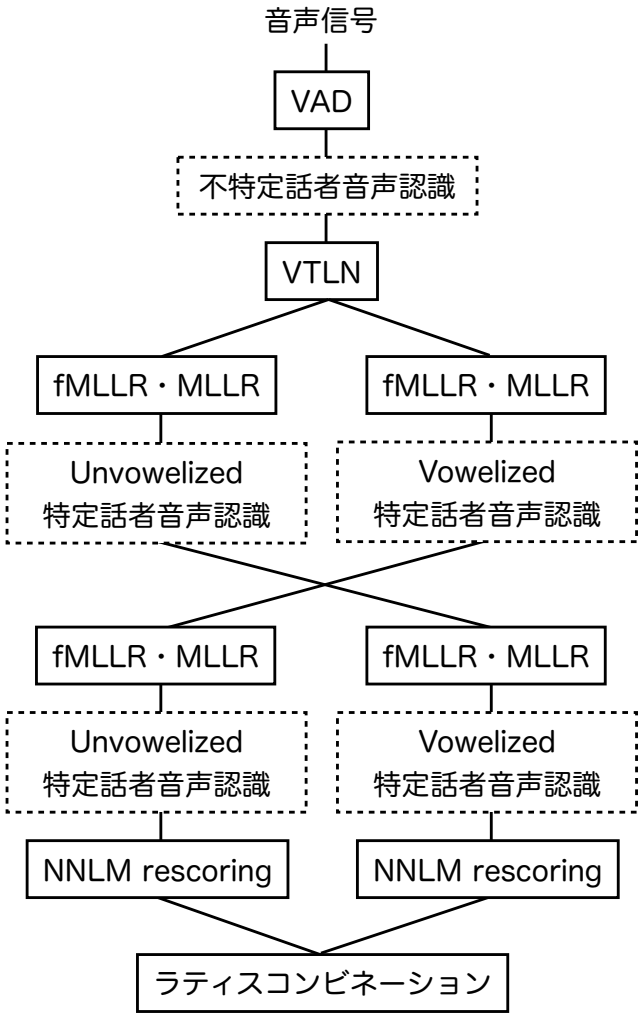

図 2.2: IBM のアラビア語音声認識システムの概要図

IBM のシステムの概要図を図 2.2 に示す．まず入力の音声信号に対し VAD を行い，まず不特定話者音声認識を行う．その際，特徴量には 13 次元の PLP を抽出し，話者毎に CMN を行う．そしてそれを 9 フレーム分連結したベクトルを LDA で 40 次元に圧縮し，それに STC をかけたものを特徴量とする．音響モデルとは，pentaphone を決定木でクラスタリングしたもの毎に用意した left-to-right 型 3 状態の HMM/GMM で，パラメタは最尤推定されている．言語モデルは，単語 4-gram を modified Kneser-Ney smoothing したものである．

次に上記の不特定話者音声認識の結果を利用して，特徴量に対し，VTLN，fMLLR と MLLR をかける．このとき，アラビア語では子音が母音化（vowelize）する現象が発生す

るため，vowelize 処理を行った場合と，行わない場合の二種類で，fMLLR/MLLR は別々に行う．そしてそれぞれで，特定話者音声認識を行う．特徴量としては先と同様の特徴量で，CMN の代わりに CMVN をかけたものを利用する．さらに，fMPE で特徴量を識別学習し，HMM/GMM のパラメタは BMMI 基準で識別学習する．

さらに上記の特定話者音声認識結果を利用して，再び VTLN 後の特徴量に対し fMLLR と MLLR をかける．このとき，unvowelized で認識したものは vowelized に，vowelized で認識したものは unvowelized にする．これにより，fMLLR/MLLR が誤りに過適応するのを防ぐ．そしてそれを先の特定話者音声認識と同様のシステムで認識を行う．

最後に，それらの認識結果を，NNLM を用いてリスコアリングし，さらにそれぞれで得られたラティスを組み合わせて，最終的な認識結果を出力する．

以上で説明した設定は，実行時間が音声信号データに対して何倍になるかを示すリアルタイムファクター（RealTime Factor; RTF）がどこまで大きくできるかによって，さまざまな設定で実験が行われている．それぞれの処理が，それぞれ少しずつ単語誤り率（Word Error Rate;WER）を低下させており，最初のシンプルな不特定話者音声認識では WER が 22.4 % だったのが，一回目の unvowelized/vowelized 特定話者音声認識で 10.1/9.4 %，二回目の特定話者音声認識で 8.7/8.6 %，NNLM resocring により 8.5/8.0 %，ラティスコンビネーションで 8.0 %，という結果が得られている．また，特徴量空間の fMPE，音響モデルの BMMI に関しては，別の設定の実験で，特徴量空間の識別学習なしで音響モデルを ML 学習したものが 17.1 %，fMPE+BMMI で 12.4 %，という結果が示されている．

また今回のタスクでは，音声に雑音が乗っていないと仮定しているので，雑音に関する処理は行われていない．また，SAT は導入されていない．

2.3.2　京都大学と NTT の衆議院会議録作成支援システム

衆議院の会議において，議事録を作成支援システムとして，京都大学と NTT が作成した音声認識システムが利用されており，これを紹介する [53].

音声信号は議長と答弁者のマイクが混ざったものと，質問者用のマイクの 2 チャネルで収音される．まず，この二つのチャネルから，どちらのチャネルを音声認識すべきかを，周波数 bin 毎のパワーを用いて選択する．次に，音響モデルが SAT されているので，VAD に加え，話者が切り替わるタイミングも適切に見つける．特徴量は MFCC + Δ + $\Delta\Delta$ で，特徴量正規化には，CMN と CMVN を行う．また VTLN も行うが，ウォーピングパラメタの推定のために，非常に簡易的なシステムを用意して，高速でパラメタを推定する．さらに，音響モデルは，教師なしで MLLR 適応する [54].

音響モデルには，HMM/GMM を MPE 基準で学習している．特徴量の識別学習は導入されていない [55]．音響モデルの学習には，学習用の音声データと，その書き起こしが必要となるが，議事録は発言の忠実な書き起こしではなく，ある程度整形された文章になって

しまっている．そこでこの整形された文章を，統計定期に話し言葉に変換し，これを使って準教師ありで音響モデルを学習している．言語モデルには，単語 3-gram で，Witten Bell スムージングが利用されている [56].

デコーダには，NTT が作成した高速な WFST デコーダを利用している．このシステムは，WFST を高速に on-the-fly で合成する手法が実装されており，これにより RTF が 1 を切るシステムが実現されている [57].

衆議院の議論でも，雑音はあまりないと考えられるため，雑音に関する処理は行われていない．

2.3.3　State-of-the-art 音声認識システム

対象となるデータや計算時間の制限など様々な要因があるのだが，現在の一つの state-of-the-art と言えるであろう音声認識システムに，[44] がある．このシステムは，Switchboard コーパスの rt03-FSH セットで，識別的言語モデルやシステムコンビネーションなどといった後段処理を行わずに 16.4% の WER を実現している．

音響モデルは HMM/DNN で，DBN を初期値として，音声認識誤りのベイズリスク最小化基準でバックプロパゲーション法で学習する．単純な方法では HMM/DNN の学習の計算量が爆発してしまうため，セカンドオーダーの最適化を並列化することで，データの増加にスケールするアルゴリズムが用いられる．言語モデルには，単語 4-gram を modified Kneser-Ney smoothing したものが用いられる．

特徴量としては，PLP 13 次元を発声単位で CMVN して，VTLN を行い，それを前後 9 フレーム連結したものを LDA して 40 次元に圧縮し，STC を行い，さらに fMLLR を行ったものが用いられる．fMMI などの特徴量空間の識別学習は，HMM/DNN を音響モデルとして利用する場合には精度の向上が得られなかったために，利用しなかったようである．

2.4　注目すべき点

2.4.1　非定常雑音に頑健な特徴量強調

最近の音声認識に関する研究の中で最も大きな出来事は，HMM/DNN を音響モデルに用いることで，非常に高い性能が実現できることが発明されたことである．HMM/DNN を使うと，音響モデルの話者適応や雑音適応は利用できないにもかかわらず，HMM/GMM 以上の性能を実現している．

音響モデルに HMM/DNN を使うとなるとすると，話者や雑音のミスマッチ問題は，モデル適応ではなく，特徴量側で解決していくことが必要になると考えられる．そのため，話

者のミスマッチに関しては，VTLN や fMLLR，雑音のミスマッチに関しては，VTS 強調や SPLICE などが有望である．

本論文では，雑音のミスマッチを特徴量側で解決する，特徴量強調法に注目する．特徴量強調では，VTS 強調や SPLICE が精度が高い手法として知られているが，それぞれに関して解決すべき問題点が残されている．

VTS 系の特徴量強調アルゴリズムは，クリーン音声 GMM のインデックスの事後確率を求める際に，分散共分散行列の逆行列を求める必要があるが，FBANK を利用する場合は対角になるため計算量が問題にならないが，MFCC を利用する場合には全角になるため，計算量がかかる．この処理は，雑音モデルが変化する度に必要になるため，非定常雑音環境下で雑音モデルが時間と共にすばやく変動する場合には現実的でなくなってしまう．結局，MFCC より精度の低い FBANK 領域を用いるか，精度の高い MFCC を使う代わりに雑音モデルが数秒の間固定したままにするか，のどちらかが必要になる．また VTS 系の特徴量強調では，特徴量として PLP や，前後数フレームの特徴量に LDA をかけた特徴量空間では利用できないことも問題点の一つである．

SPLICE は，任意の特徴量空間で利用することができて，しかも非常に高速に動作する．しかし，ステレオデータを用いる手法であるため，突発的な非定常雑音など，学習用ステレオデータの雑音環境に含まれていない雑音が重畳してしまった場合には，正しく特徴量強調を行うことができない．その一つの解決策として NMN-SPLICE があるが，NMN-SPLICE 対数をとった後の特徴量空間において引き算を行うというヒューリスティックな手法であり，なぜそれでうまく動作するのかには疑問が残る．

本論文では，高速に動作するために実用的であるという点で，SPLICE などのステレオデータを用いる特徴量強調に注目する．そして，ステレオベースの特徴量強調を非定常雑音にも頑健になるように改良する手法を提案する．具体的には，区分的線形変換において各部分空間の事後確率を求める部分の計算を，クリーン音声状態の識別と捉える考え方を導入し，その入力特徴量として，観測したノイジー音声の特徴量に加え，推定した雑音特徴量や，前後数フレームの特徴量を入力として利用することを提案する．

第 3 章 において，この手法について詳しく述べる．

2.4.2　ミスマッチのない場合にも頑健な雑音抑圧

雑音抑圧技術を利用すると，雑音が含まれたノイジーな音声の認識精度は向上する．しかし，クリーン音声に対して雑音抑圧を行うと，雑音抑圧を行わなかった場合と比較して，精度が低下してしまう場合もある．

雑音抑圧手法の一つである DNA では，DNA-CD と呼ばれる技術を用いて，入力がクリーン音声らしければ雑音抑圧を行わず，入力がノイジー音声らしければ雑音抑圧を行う，ということが行われている．このような技術は，DNA に限らず，あらゆる雑音抑圧手法

において有効だと考えられる．

そこで本論文では，あらゆる雑音抑圧手法において利用できる技術として，予めクリーン音声・ノイジー音声を識別し，その結果に依存して雑音抑圧を行うか行わないかを決定する手法を提案する．

第 4 章 において，この手法について詳しく述べる．

2.4.3　識別的リランキングにおける音声の構造的表象の利用

音響モデルの研究からでてきた識別モデルを用いた音声認識と，識別的言語モデルの研究は，ほぼ同じような手法と目的を持ちつつ，ここまで互いに独立に発展してきている．

本論文で特に注目するのは，音響モデル側の研究では識別モデルを使うことそのものに注目した研究が多く，どのような特徴量を用いるか，特に長い時間範囲にまたがる特徴量の利用について，ないがしろにされていた点である．逆に識別的言語モデルの研究では，NNLM の尤度など，文全体にまたがる特徴量を積極的に利用しようとする研究が行われている．

そこで本論文では，識別的言語モデルで広く用いられている N-best リストの識別的リランキング手法において，長時間にわたって定義される音響的特徴量を利用する手法を提案する．この情報は，これまで利用されていなかった側面の情報であるため，state-of-the-art 音声認識システムの認識精度をさらに向上させられる可能性がある．具体的には，この長時間にわたる音響特徴量として，音声の構造的表象を利用する [58]．音声の構造的表象とは，話者の違いに非常に高い頑健性を持つ特徴で，これまで孤立単語音声認識 [59] や外国語自動発音評価 [60] に利用され，効果が示されている．本論文の提案手法は，音声の構造的表象を初めて大語彙音声認識に適用する手法となる．

第 5 章において，この手法について詳しく述べる．

第3章

非定常雑音に頑健なステレオベース特徴量強調

3.1　はじめに

スマートフォンでの音声認識など，実環境下で音声認識を利用する機会が増えている．しかし実環境下では，定常・非定常の背景雑音により音声が劣化するため，認識性能が低下してしまう．そのため，実環境で用いられる音声認識システムは，耐雑音性を備えることが望ましい．

一つのマイクで収音された音声のの耐雑音音声認識は，二つのアプローチに分類できる．一つ目は音響モデルである HMM/GMM を雑音環境に適応させるモデル適応，二つ目は入力音声から雑音成分を抑圧する雑音抑圧である．前者としては，PMC や VTS 適応を用いたモデル適応法が有名である．これらの手法は，高い精度が得られるが，音響モデルのパラメータを適応するために，相当の計算コストがかかる．そのため，非定常雑音環境下において，雑音の動きに追随して適応を行うことは現実的に難しい．また，音響モデルとして HMM/DNN を用いる場合には利用できないという問題もある．そこで，非定常雑音環境下でも頑健に音声認識を行うために，後者の雑音除去の方に注目する．

雑音除去は，さらに二つのアプローチに分類できる．一つ目はパワースペクトルドメインで雑音除去する方法，二つ目は音声認識に用いる特徴量ドメイン（MFCC + Δ + ΔΔ など）で雑音除去を行う特徴量強調である．前者としては，SS などがあるが，雑音除去後のスペクトル歪みが大きくなってしまう問題がある．後者としては，SPLICE や VTS 強調などがある．これらは，クリーン音声の特徴量に関する情報を事前に学習するため，音声認識に適した雑音除去が実現できる．

近年成功している特徴量強調のほとんどは，ノイジー音声の特徴量の分布を GMM でモデル化して，それを利用した区分的な変換により，特徴量強調を行っている．具体的には，ノイジーな音声が観測されると，GMM インデックスの事後確率を求め，この事後確率で，各インデックス毎の変換処理を重み付けして利用する．例えば VTS 強調では，事前に用意したクリーン音声の GMM と推定した雑音パラメータから，VTS 近似を用いることでノイジー音声の GMM を合成し，この事後確率を元に変換処理を行う．ノイジー音声の GMM を雑音パラメータを用いて合成することで，GMM を処理対象の雑音環境に適合させることができる．そのため，精度が高くなるが，特に特徴量として MFCC +Δ + ΔΔ を利用した場合実用的な計算コストで処理できなくなってしまう欠点がある．

一方 SPLICE では，クリーン音声とノイジー音声のステレオデータを学習データ用い，ノイジー音声に関する GMM を用意してそのインデックスの事後確率を元に区分的線形変換を学習して利用する．これにより特徴量強調時の計算量は非常に低くなるが，GMM の学習データが必ずしも処理対象の雑音環境に適合しないため，GMM による近似が不適切になる場合が多く，精度面で劣る．

これを鑑みて本稿では，ノイジー音声状態の事後確率の計算を，クリーン音声状態の事後確率の計算と近似し，これを識別的に推定する手法を提案する．そしてその事後確率を

元に，SPLICE と同様，区分的な線形変換を予め学習して利用する．これにより，計算量を低く保ったまま，所望の事後確率を高精度で推定することが可能になる．

また，従来の VTS 強調や SPLICE では，ある時間フレーム t のクリーン音声特徴量を推定するために，その時間 t におけるノイジー音声特徴量や，雑音の分布パラメタのみを利用している．しかし，LDA や fMPE/fBMMI などの特徴量の識別学習やタンデムアプローチなどにおいて，前後数フレーム分の特徴量を利用することで精度が向上した結果から考えると，特徴量強調においても，前後数フレーム分の特徴量を利用することが効果的かもしれない．

前後数フレーム分の特徴量の利用は，GMM のインデックスの事後確率を求めるときだけでなく，インデックス毎の変換処理においても利用可能できる．提案手法では，SPLICE と同じく，線形変換を予め学習して利用するが，これを前後数フレーム分の特徴量を連結した特徴量の線形変換に置き換えることで，さらに精度を高める手法を提案する．この際，次元が高くなり過学習が発生するので，正則化を導入する．

提案手法は，ステレオデータを学習データとして利用するので，競合相手は SPLICE やそれを改良した NMN-SPLICE である．提案手法を AURORA2 データベースで評価したところ，それらと比較して有意に WER が低くなることが分かった．

3.2　従来法の解釈

3.2.1　クリーン音声状態推定としての VTS 強調の解釈

クリーン音声特徴量を $\boldsymbol{x}_t$，ノイジー音声特徴量を $\boldsymbol{y}_t$，雑音特徴量の分布を $p(\boldsymbol{n}_t) = \mathcal{N}(\boldsymbol{n}_t; \boldsymbol{\mu}^{\boldsymbol{n}_t}, \boldsymbol{\sigma}^{\boldsymbol{n}_t})$ とおく．第 1 章では，雑音特徴量の分布は時間によらないと仮定していたが，ここでは非定常な雑音を想定して，雑音の分布も時間インデックス t によって変動しているとする．

VTS 強調では，まず以下のように $\boldsymbol{x}_t$ が時間に依らず GMM から出力されていると仮定して，以下のような混合数 K の GMM を学習する．

$$p(\boldsymbol{x}_t) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_k^{\boldsymbol{x}}, \boldsymbol{\sigma}_k^{\boldsymbol{x}}) \tag{3.1}$$

この GMM から，$\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{n}_t$ の関係を VTS 近似することで，以下のように $\boldsymbol{y}_t$ の確率密度

関数を GMM で合成する．

$$p(\boldsymbol{y}_t) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_k^{\boldsymbol{y}}, \boldsymbol{\sigma}_k^{\boldsymbol{y}}) \tag{3.2}$$

$$\boldsymbol{\mu}_k^{\boldsymbol{y}} = g(\boldsymbol{\mu}_{k_t}^{\boldsymbol{x}}, \boldsymbol{\mu}^{\boldsymbol{n}_t}) \tag{3.3}$$

$$\boldsymbol{\sigma}_k^{\boldsymbol{y}} = \frac{\boldsymbol{\sigma}_{k_t}^{\boldsymbol{x}}}{(\mathbf{1} + \exp(\boldsymbol{\mu}^{\boldsymbol{n}} - \boldsymbol{\mu}_{k_t}^{\boldsymbol{x}}))^2} + \frac{\boldsymbol{\sigma}^{\boldsymbol{n}}}{(\exp(\boldsymbol{\mu}_{k_t} - \boldsymbol{\mu}^{\boldsymbol{n}}) + \mathbf{1})^2} \tag{3.4}$$

ただし，この式は (2.12) から得られる式で，特徴量は FBANK を仮定している．特徴量が変化した場合も，この計算式を変更するだけで，形式自体は同じである．

この合成されたノイジー音声特徴量の GMM を利用し，VTS 強調では $p(k|\boldsymbol{y}_t)$ を計算し，特徴量強調に利用する．ここで，$p(k|\boldsymbol{y}_t)$ は理想的にはクリーン音声特徴量の GMM から計算できる $p(k|\boldsymbol{x}_t)$ を近似しようとしている（実際の特徴量強調時には $\boldsymbol{x}_t$ は既知ではないので $p(k|\boldsymbol{x}_t)$ は計算できない）．

ここで，インデックス k は，音素の種類や音色の違いなどといった,「クリーン音声状態」を表していると解釈できる．すなわち，$p(k|\boldsymbol{x}_t)$ は，どのようなクリーン音声状態をとるかの事後確率であり，VTS 強調ではそれを $p(k|\boldsymbol{y}_t)$ で近似していると解釈できる．

さらに言うと，一般的に GMM のインデックスの事後確率は，ある k において 1 に近い値を取り，それ以外ではほぼ 0 となるような，スパースな状態になる．すなわち $\{p(k|\boldsymbol{x}_t)\}_{k=1\cdots K}$ の情報は，

$$k_t^* = \operatorname*{argmax}_k p(k|\boldsymbol{x}_t) \tag{3.5}$$

なる k_t^* の情報とほぼ同値である．k_t^* は，まさにある時刻 t における「クリーン音声状態」を表している．すなわち VTS 強調は，クリーン音声状態を推定し，それを元に強調処理を行っていると解釈できる．

3.2.2　ノイジー音声状態推定としての SPLICE の解釈

SPLICE による特徴量強調では，まず学習データを使って $\boldsymbol{y}_t$ を出力する混合数 K の GMM を学習する．

$$p(\boldsymbol{y}_t) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_k^{\boldsymbol{y}}, \boldsymbol{\sigma}_k^{\boldsymbol{y}}) \tag{3.6}$$

強調時には，これを用いて $p(k|\boldsymbol{y}_t)$ を計算して利用する．

ここで，インデックス k は，音素の種類や音色に加え，雑音の種類等を含む,「ノイジー音声状態」を表していると解釈できる．そしてここでいう「ノイジー音声」とは，学習データに含まれる雑音環境のみが考慮されている．

この意味で，VTS 強調における $p(k|\boldsymbol{y}_t)$ の計算と SPLICE における $p(k|\boldsymbol{y}_t)$ の計算は，大きく意味が異なる．VTS 強調ではクリーン音声状態を推定しようとしているため，どの

ような雑音環境にも対処できる（ただし計算量は高い）．SPLICE ではノイジー音声状態を推定しているため，学習データに含まれない雑音環境は考慮できない（ただし計算量は低い）．

NMN-SPLICE では，$\boldsymbol{y}_t$ の代わりに，$\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$ を利用する．ただし $\hat{\boldsymbol{n}}_t$ は，時刻 t における雑音の特徴量である．NMN-SPLICE では，ノイジー音声 $\boldsymbol{y}_t$ の状態の代わりに，$\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$ の状態を推定して利用することになる．ここで，音声が存在しない無音区間では，$\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t = \boldsymbol{0}$ が近似的に成り立つため，$\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$ の空間は，$\boldsymbol{y}_t$ の空間と比べて「汚れていない」空間だと考えることができる．すなわち，$\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$ 状態も音声や話者に加え雑音の情報も持っているが，$\boldsymbol{y}_t$ 状態と比較すると，特に無音区間において，雑音の情報が少ないと考えられる．この意味で，NMN-SPLICE は SPLICE と比べて雑音環境のミスマッチに強いと考えられる．ただし，$\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$ の空間が最も優れているという保障はなく，NMN-SPLICE ヒューリスティックなやり方である．

3.3　提案手法

3.3.1　クリーン音声状態識別に基づく特徴量強調

提案手法では，ステレオデータを学習データとして，VTS 近似を用いずにクリーン音声状態 k^* を識別し，それを用いて特徴量強調を行う．提案手法は，VTS 近似を用いず，ステレオデータから k^* 識別関数を学習するため，処理が非常に高速である．また NMN-SPLICE では，$\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$ の GMM を学習し利用するという，ヒューリスティックなやり方が採用されているが，提案手法ではこれをクリーン音声状態を識別するという明確な目的に置き換える．これにより，NMN-SPLICE を越える精度を目指す．

提案手法ではまず，以下のようにクリーン音声特徴量の GMM を学習する．

$$p(\boldsymbol{x}_t) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_k^{\boldsymbol{x}}, \boldsymbol{\sigma}_k^{\boldsymbol{x}}) \tag{3.7}$$

そして，$\{p(k|\boldsymbol{x}_t)\}_{k=1\cdots K}$ もしくは $k_t^* = \mathrm{argmax}_k\, p(k|\boldsymbol{x}_t)$ を，$\boldsymbol{x}_t$ を観測することなく，推定することが最初の目的である．その際利用する特徴量は，ノイジー音声特徴量 $\boldsymbol{y}_t$ と，雑音特徴量の推定値 $\hat{\boldsymbol{n}}_t$ と，それらの前後数フレーム分の特徴量である．

そこで，k_t^* の推定値 $\hat{k}_t^*$ を推定するために，以下のような識別モデルを考える．

$$\hat{k}_t^* = \underset{k}{\operatorname{argmax}} f(k_t^* = k|\boldsymbol{d}_t;\theta) \tag{3.8}$$

$$\boldsymbol{d}_t = \begin{bmatrix} \vdots \\ \boldsymbol{y}_{t-1} \\ \hat{\boldsymbol{n}}_{t-1} \\ \boldsymbol{y}_t \\ \hat{\boldsymbol{n}}_t \\ \boldsymbol{y}_{t+1} \\ \hat{\boldsymbol{n}}_{t+1} \\ \vdots \end{bmatrix} \tag{3.9}$$

ただし θ は識別モデルのパラメータである．また $\boldsymbol{d}_t$ は，前後あわせて 9 フレーム程度のノイジー音声特徴量，雑音特徴量の推定値を連結したベクトルである．

識別モデルには，任意の識別モデルを用いることができる．例えば SVM（Support Vector Machine）や，ロジスティック回帰，ニューラルネットワークなど，また時系列の識別問題なので CRF（Conditional Random Field）なども利用できる．また他に，LDA などで識別的に特徴量空間を次元圧縮し，その空間で新たに GMM を学習してそのインデックスの事後確率を利用するように式を変えてもよい．識別モデルのパラメタ θ の学習データには，ステレオデータを利用する．$\boldsymbol{x}_t$ が得られれば，対応する k_t^* が (3.7) から求められるので，それと $\boldsymbol{d}_t$ を使えば学習が行える．

ここで，k_i^* を求める識別モデルを利用することは，LDA を用いた特徴量空間の識別学習や，タンデムアプローチなどと関連深いことを指摘しておく．LDA やタンデムアプローチでは，強制アライメントによって得られた HMM の状態ラベルを推定するように，LDA あるいはニューラルネットワークなどを学習している．ここで提案手法で利用しているラベル k_i^* も，GMM によるクリーン音声状態のラベルであるため，用いているラベルは似たような情報を持っているといえる．そのため，提案手法によるクリーン音声状態識別は，LDA やタンデムアプローチと，目的は異なるものの，処理としては似ている処理を行なっていると言うことができる．

3.3.2　ソフトな LDA を用いた手法

提案手法によるクリーン音声状態識別では，任意の識別モデルを利用することができるが，ここでは，予備実験の結果最も精度が高くなった，ソフトな LDA を用いる手法について詳細を述べる．ここで「ソフトな」とは，教師ラベルとして k_t^* ではなく，$\{p(k|\boldsymbol{x}_t)\}_{k=1\cdots K}$ の形のまま利用するという意味と，最終的な結果が，$\hat{k}_t^*$ としてでなく，$\{p(s|\boldsymbol{d}_t\}_{s=1\cdots S}$ の

ような形で得られる，という二つの意味がある．

まず，学習データには，$\{\{p(k|\boldsymbol{x}_t)\}_{k=1\cdots K}, \boldsymbol{d}_t\}_{t=1\cdots T}$ を用いて，以下のように $\boldsymbol{L}\boldsymbol{d}_t$ のように次元圧縮を行う行列 $\boldsymbol{L}$ を求める．

$$\boldsymbol{L} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \frac{\boldsymbol{W}^\top \boldsymbol{\Sigma}^w \boldsymbol{W}}{\boldsymbol{W}^\top \boldsymbol{\Sigma}^b \boldsymbol{W}} \tag{3.10}$$

$$\boldsymbol{\Sigma}^w = \sum_{k=1}^{K}\sum_{t=1}^{T} p(k|\boldsymbol{x}_t)\left(\boldsymbol{d}_t - \boldsymbol{\mu}_k^w\right)\left(\boldsymbol{d}_t - \boldsymbol{\mu}_k^w\right)^\top \tag{3.11}$$

$$\boldsymbol{\Sigma}^b = \sum_{k=1}^{K}\left(\sum_{t=1}^{T} p(k|\boldsymbol{x}_t)\right)\left(\boldsymbol{\mu}_k^w - \frac{\sum_{t=1}^{T}\boldsymbol{d}_t}{T}\right)\left(\boldsymbol{\mu}_k^w - \frac{\sum_{t=1}^{T}\boldsymbol{d}_t}{T}\right)^\top \tag{3.12}$$

$$\boldsymbol{\mu}_k^w = \frac{1}{\sum_{t=1}^{T} p(k|\boldsymbol{x}_t)}\sum_{t=1}^{T} p(k|\boldsymbol{x}_t)\boldsymbol{d}_t \tag{3.13}$$

ここで，クラス内共分散行列 $\boldsymbol{\Sigma}^w$ が，$p(k|\boldsymbol{x}_t)$ の重み付け和で表現されているところが，通常の LDA と異なる部分である．(3.10) の解析解は，$(\boldsymbol{\Sigma}^w)^{-1}\boldsymbol{\Sigma}^b$ の固有ベクトルを固有値が小さい順に並べることで得ることが出来る．

次に，次元圧縮された空間 $\boldsymbol{v}_t = \boldsymbol{L}\boldsymbol{d}_t$ において，混合数 S の GMM を以下のように学習する．

$$p(\boldsymbol{v}_t) = \sum_{s=1}^{S} \pi_s^{\boldsymbol{v}} \mathcal{N}(\boldsymbol{v}_i; \boldsymbol{\mu}_s^{\boldsymbol{v}}, \boldsymbol{\sigma}_s^{\boldsymbol{v}}) \tag{3.14}$$

ここで $\pi_s, \boldsymbol{\mu}_s^{\boldsymbol{v}}, \boldsymbol{\Sigma}_s^{\boldsymbol{v}}$ は，それぞれ s 番目のインデックスに対応 s るう GMM の重み，平均ベクトル，分散ベクトルである．

ここで，$\boldsymbol{v}_t$ は，$\{p(k|\boldsymbol{x}_t)\}_{k=1\cdots K}$ の情報を保存するように次元圧縮された空間であるので，その空間内で学習された GMM のインデックスの事後確率 $\{p(s|\boldsymbol{v}_t)\}_{s=1\cdots S}$ は，$\{p(k|\boldsymbol{x}_t)\}_{k=1\cdots K}$ の k とは直接的関係はないものの，似たような情報を持っていると考えられる．特徴量強調の事を考えると，クリーン状態が異なる場合に異なるインデックスの事後確率が高くなればよいので，$\{p(s|\boldsymbol{v}_t)\}_{s=1\cdots S}$ を直接特徴量強調に利用できる．すなわち，以下のように観測された特徴量 $\boldsymbol{d}_t$ からインデックス s の事後確率を計算する．

$$p(s|\boldsymbol{d}_t) = \frac{\pi_s^{\boldsymbol{v}} \mathcal{N}(\boldsymbol{v}_t; \boldsymbol{\mu}_s^{\boldsymbol{v}}, \boldsymbol{\Sigma}_s^{\boldsymbol{v}})}{\sum_{s=1}^{S} \pi_s^{\boldsymbol{v}} \mathcal{N}(\boldsymbol{v}_t; \boldsymbol{\mu}_s^{\boldsymbol{v}}, \boldsymbol{\Sigma}_s^{\boldsymbol{v}})} \tag{3.15}$$

3.3.3　インデックス毎の線形変換

$\hat{k}_t^*$ を識別モデルで求めた後は，以下のような線形変換でクリーン音声特徴量の推定値を求める．

$$\hat{\boldsymbol{x}}_t = \boldsymbol{A}_{\hat{k}_t^*}\boldsymbol{e}_t \tag{3.16}$$

ソフトな LDA を用いて $\{p(s|\boldsymbol{d}_t)\}_{s=1\cdots S}$ を求めた場合には，以下のような区分的線形変換で求める．

$$\hat{\boldsymbol{x}}_t = \sum_{s=1}^{S} p(s|\boldsymbol{d}_t)\boldsymbol{A}_s\boldsymbol{e}_t \tag{3.17}$$

ただしここで $\boldsymbol{e}_t$ は，$\boldsymbol{d}_t$ と同じようにノイジー音声特徴量と雑音特徴量の推定値の前後数フレームを結合し，さらにバイアス項のために 1 を結合した以下のような特徴量である．

$$\boldsymbol{e}_t = \begin{bmatrix} 1 \\ \vdots \\ \boldsymbol{y}_{t-1} \\ \hat{\boldsymbol{n}}_{t-1} \\ \boldsymbol{y}_t \\ \hat{\boldsymbol{n}}_t \\ \boldsymbol{y}_{t+1} \\ \hat{\boldsymbol{n}}_{t+1} \\ \vdots \end{bmatrix} \tag{3.18}$$

線形変換 $\{\boldsymbol{A}_k\}_{k=1\cdots K}$ もしくは $\{\boldsymbol{A}_s\}_{s=1\cdots S}$ は，SPLICE と同様に，ステレオデータを用い重み付き最小二乗誤差基準で推定することが出来る．ここで，SPLICE や NMN-SPLICE では，$\boldsymbol{y}_t$ もしくは $\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$ の線形変換を考えていたので，$\boldsymbol{A}$ は，特徴量の次元数を M として，$M \times (M+1)$ 行列であった（+1 はバイアス項に相当する）．しかし，今回のように線形変換の入力として $\boldsymbol{e}_t$ を用いると，前後 9 フレーム分の特徴量を使う場合には $\boldsymbol{A}$ が $M \times (18M+1)$ 行列となり，求めるべきパラメタ数が非常に沢山になってしまい，過学習の問題が発生してしまう．

そこで，重み付き最小二乗誤差基準に，$\boldsymbol{A}$ の各要素が小さくなるように正則化を行い，過学習を抑制する．二次の正則化項を導入する場合の $\{\boldsymbol{A}_s\}_{s=1\cdots S}$ の学習は，以下のようにかける．

$$\boldsymbol{A}_s = \underset{\boldsymbol{A}_s}{\operatorname{argmin}} \sum_{t=1}^{T} p(s|\boldsymbol{d}_t)\,||\boldsymbol{x}_t - \boldsymbol{A}_s\boldsymbol{e}_t||^2 - \lambda R_s \tag{3.19}$$

$$R_s = \frac{\mathbf{1}^\top {\boldsymbol{A}_s}'^\top \mathrm{diag}(\boldsymbol{E}\boldsymbol{P}_s\boldsymbol{E}^\top)\boldsymbol{A}'_s\mathbf{1}}{M} \tag{3.20}$$

$$\boldsymbol{E} = [\boldsymbol{e}_1, \cdots, \boldsymbol{e}_T] \tag{3.21}$$

$$\boldsymbol{P}_s = \begin{pmatrix} p(s|\boldsymbol{d}_1) & 0 & \cdots & 0 \\ 0 & p(s|\boldsymbol{d}_2) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & p(s|\boldsymbol{d}_T) \end{pmatrix} \tag{3.22}$$

ここで (3.19) の第二項が，L2 正則化項に相当する．$\boldsymbol{A}_s{}'$ とは，$\boldsymbol{A}_s$ のバイアス項に相当する部分を取り除いた $M \times 18M$ 行列である．

ここで，SPLICE の線形変換では，バイアス項の成分が重要で，それ以外は 0 と仮定しても精度は大きく低下しないという知見がある [22]．今回の手法では，クリーン音声状態を識別するが，クリーン音声状態が完璧に分かれば，その状態に対応する代表的なベクトルを出力するだけでも十分な特徴量強調が実現できると考えられるので，提案手法においてもバイアス項成分は非常に重要だと考えられる．

そこで正則化項には，$\boldsymbol{A}_s{}'$ を使い，バイアス項に関しては正則化をかけず，それ以外のパラメタに関してのみ，値が 0 に近くなるように正則化をかけることにする．λ は，正則化の強さを決めるパラメタで，予め適当に設定する．$\mathrm{diag}(\boldsymbol{D}'\boldsymbol{P}_s\boldsymbol{D}'^{\top})$ という項は，各次元間でのオーダーの違いの影響を消すための項である．

(3.19) は解析的に解くことができ，その解は以下のように求められる．

$$\boldsymbol{A}_s = \boldsymbol{X}\boldsymbol{P}_s\boldsymbol{E}^{\top}(\boldsymbol{E}\boldsymbol{P}_s\boldsymbol{E}^{\top} + \lambda\boldsymbol{I}'\mathrm{diag}(\boldsymbol{E}\boldsymbol{P}_s\boldsymbol{E}^{\top}))^{-1}, \tag{3.23}$$

ここで $\boldsymbol{X}, \boldsymbol{E}$ は $\boldsymbol{x}_t$ や $\boldsymbol{e}_t$ を並べた行列，$\boldsymbol{P}_s$ は $p(s|\boldsymbol{d}_t)$ を対角成分にもつ対角行列，$\boldsymbol{I}'$ は，$(18M+1)\times(18M+1)$ 行列で，$I'_{1,1}=0$ となり，それ以外の対角要素は 1，それ以外はすべて 0 となるような行列である．

3.4　実験

3.4.1　データベース

実験に用いるデータベースには，AURORA2 データベースを利用する [61]．

AURORA2 データベースは，雑音環境下における連続数字音声認識の評価を行うデータベースである．学習用のデータとして，成人男性 55 名，成人女性 55 名 による合計 8440 発声のクリーンな連続数字読み上げ音声が与えられる．これを clean データと呼ぶ．またそれらを 422 発声ずつに 20 分割し，4 種類（Subway, Babble, Car, Exhibit）× 5 SNR (5, 10, 15, 20, ∞ [dB]) の合計 20 環境の加法性雑音を重畳されたデータも与えられる．これを multi データと呼ぶ．

評価用のデータは，A セット，B セット，C セットの 3 つに分かれている．音声データとしては，成人男性 52 人，成人女性 52 人による合計 4004 発声文の音声があり，それを 4 つに分割した 1001 発声が基本単位となっている．A セットでは，学習用の multi データに含まれる雑音と同じ種類の雑音（Subway, Babble, Car, Exhibit）が，4 つの分割されたデータにそれぞれ重畳されている．SNR は -5, 0, 5, 10, 15, 20, ∞ [dB] の 7 つが用意されている．A セットを用いれば，雑音環境クローズドの場合の評価が行える．

B セットには，学習データとは異なる 4 種類の雑音（Restaurant, Street, Airport, Station）

が重畳されている．SNR は同様の 7 つが用意されている．B セットを用いれば，雑音環境オープンの場合の実験を行うことができる．

C セットは，A セットの Subway 雑音，B セットの Street データに対して，さらに電話通信を想定したフィルタを通して乗法性雑音を加えたものである．C セットを用いれば，フィルタがかかった場合の評価が行える．

なお最終的な手法の優劣を見るために，各セットの WER の平均をとるときには，7 種類すべての SNR を用いず，0, 5, 10, 15, 20 [dB] の 5 つの平均を見ることが多いため，本論文でもそれに沿うことにした．

3.4.2　特徴量強調法に関する実験

まず，クリーン音声状態識別と前後数フレーム分のノイジー音声・雑音特徴量を利用をした特徴量強調について，以下のことを実験的に調べる．

- SPLICE，NMN-SPLICE と比べて提案手法の優位性はあるのか．
- クリーン音声状態の識別，雑音特徴量の利用，前後数フレーム分の特徴量利用，のどの部分が精度に大きな影響を与えるのか．
- 正則化パラメタ λ の変化でどの程度精度が影響を受けるのか．

音響モデルの学習には，他の文献と比較しやすくするため，complex backend と呼ばれる AURORA2 の評価で標準的に用いられている設定を利用した．ただし，多くの文献行われているように $\boldsymbol{y}$，$\boldsymbol{n}$，$\hat{\boldsymbol{x}}$ の特徴量として，パワーの代わりに MFCC の 0 次元を利用するよう変更した．さらに，特徴量のフレーム長を 50 msec ，シフト長を 10 msec とした[1]．これら以外の設定は，すべて complex backend のものをそのまま利用した [63]．

また，学習データの clean データのみを使って学習したもの（clean 条件）と，multi データを使い，学習時にも評価時と同じ特徴量強調を行なって NAT したもの（multi 条件）の二種類の音響モデルを用意して評価した．また，SPLICE や提案手法ではステレオデータを学習データに利用するために，multi データを利用する．そのため基本的には，multi 条件の実験結果が重要である．

なお SPLICE や提案手法において，GMM と線形変換の学習に用いる multi データとしては，8440 発声をすべて使わず，クリーン音声（SNR が ∞ の場合）の音声を取り除いた 6752 発声分のみを用いた．

実験において設定した値などは以下の通りである．

- $\hat{\boldsymbol{n}}_t$ は，[62] の方法を用いて推定したものを利用する．
- クリーン音声 GMM の混合数 K は 1024 に固定．

[1]通常はフレーム長 25 msec，シフト長 10 msec が用いられることが多いが，雑音推定アルゴリズムとして用いる [62] が，フレーム長が 50 msec である場合の方が精度が高くなるためこれを採用した．

- クリーン音声状態識別のために，ソフトな LDA を導入する．
- LDA 後の特徴量空間の次元は 39 に固定．
- LDA 後の特徴量空間における GMM の混合数 S は 1024 に固定．
- SPLICE や NMN-SPLICE で利用する GMM の混合数も 1024 に固定．
- 特徴量正規化として発声単位の CMN をかける．

表 3.1: 様々な条件での AURORA2 データベースにおける WER (%) の平均．「区分」は，その特徴量の GMM を利用することを，「線形変換」はその特徴料の線形変換を利用することを意味する．また括弧内の数字は，前後何フレームの特徴量を利用したかを示す．

			clean 条件				multi 条件			
区分	線形変換	λ	A	B	C	平均	A	B	C	平均
特徴量強調なし			30.62	25.67	30.05	24.53	7.80	7.95	7.53	7.81
SPLICE		0	10.73	12.51	14.06	12.11	6.73	11.15	9.10	8.97
$\boldsymbol{y}_t$	$\boldsymbol{e}_t$ (1)	0	9.31	10.78	12.29	10.49	5.85	10.24	8.92	8.22
$\boldsymbol{y}_t$	$\boldsymbol{e}_t$ (9)	10^{-3}	8.78	9.81	11.45	9.73	**5.57**	8.91	8.74	7.54
NMN-SPLICE		0	10.17	10.08	10.45	10.19	6.61	8.29	6.66	7.29
$\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$	$\boldsymbol{y}_t$	0	10.30	10.32	11.69	10.59	7.00	8.66	7.54	7.77
$\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$	$\boldsymbol{e}_t$ (1)	0	9.36	9.46	10.40	9.61	6.46	8.13	6.56	7.15
$\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$	$\boldsymbol{e}_t$ (9)	10^{-3}	7.98	8.47	**9.22**	8.42	6.07	7.78	**6.52**	6.84
$\boldsymbol{v}_t(1)$	$\boldsymbol{y}_t$	0	8.99	9.45	12.92	9.96	6.88	8.84	9.67	8.22
$\boldsymbol{v}_t(1)$	$\boldsymbol{e}_t$ (1)	0	8.21	8.51	11.50	8.99	6.54	8.20	8.62	7.62
$\boldsymbol{v}_t(1)$	$\boldsymbol{e}_t$ (9)	10^{-3}	7.25	7.52	10.11	7.93	5.61	**7.31**	7.58	**6.68**
$\boldsymbol{v}_t(9)$	$\boldsymbol{y}_t$	0	7.81	8.20	10.81	8.57	7.21	8.71	9.75	8.32
$\boldsymbol{v}_t(9)$	$\boldsymbol{e}_t$ (1)	0	7.20	7.56	10.09	7.92	6.47	7.98	8.88	7.56
$\boldsymbol{v}_t(9)$	$\boldsymbol{e}_t$ (9)	10^{-3}	**6.50**	**7.19**	9.42	**7.36**	6.06	7.37	8.17	7.01

区分線形変換の「区分」で用いる GMM の特徴量と，「線形変換」で用いる特徴量を変えながら実験を行い，提案手法と SPLICE や NMN-SPLICE の比較，およびどの部分で違いがでるのかを調べた．結果を表 3.1 に示す[2]．正則化パラメタ λ に関しては，予備実験を行い線形変換に前後 9 フレームの特徴量を用いる場合（特徴量の次元が 703 次元）のみ，10^{-3} とし，それ以外では 0 に設定して正則化を行わない条件で実験を行うことにした．SPLICE は，「区分」に $\boldsymbol{y}_t$ の GMM，「線形変換」に $\boldsymbol{y}_t$ の線形変換を用いることに相当し，NMN-SPLICE は，「区分」に $\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$ の GMM，「線形変換」では $\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$ の線形変換で $\boldsymbol{x}_t - \hat{\boldsymbol{n}}_t$ を推定し，最後に $\hat{\boldsymbol{n}}_t$ を足す処理に相当する．

まず，SPLICE，NMN-SPLICE を他の提案手法と比べると，clean 条件でも multi 条件でも，それを越える精度が実現されている手法が存在している．そのため，提案手法によ

[2]ただし，ここではセットごとの WER の平均値のみを載せている．詳細な WER は付録にある．

り，SPLICE や NMN-SPLICE の精度を改善できることが分かる．

次に，提案手法のどの部分が精度向上に効果があるのかについて考察する．まず「区分」のやり方を固定して，線形変換で使う特徴量を $\boldsymbol{y}_t$，$\boldsymbol{e}_t(1)$，$\boldsymbol{e_t}(9)$ ていった場合を見ると，どの場合においても，$\boldsymbol{y}_t$ より $\boldsymbol{e}_t(1)$，$\boldsymbol{e}_t(1)$ より $\boldsymbol{e}_t(9)$ の性能が向上している．これにより，線形変換の特徴量としては，ノイジー音声特徴量だけでなく，雑音特徴量の推定値や，前後の特徴量を連結して用いた方が，精度が高くなることが分かった．これは，前後数フレームの特徴量を用いることにより，非定常雑音により突発的に汚れてしまった特徴量を補正できた効果であると考察できる．

線形変換に用いる特徴量を $\boldsymbol{e_t}(9)$ に固定して，区分のために用いる GMM の特徴量を変えた場合の結果を見ると，認識の条件によって傾向が異なっている．まず，評価データとして，雑音環境クローズド条件である A セット を利用し．multi 条件で学習した HMM/GMM を利用した場合は，単純な SPLICE の精度が最も高くなっている．雑音環境のミスマッチがなければ,「区分」も雑音に合わせて行った方がより適切な区分的線形変換が学習できると考えられるので，この結果は妥当である．

一方，雑音環境のミスマッチがある B セットでは，$\boldsymbol{y}_t$ の GMM より $\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$ の GMM, が精度が高く，さらに $\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$ の GMM より $\boldsymbol{v}_t(1)$ または $\boldsymbol{v}_t(9)$ の GMM を用いた方が精度が高くなっている．これにより，雑音環境のミスマッチがある場合には，提案手法によるクリーン音声状態識別を導入する効果があることが分かる．ただしこのときは，前後数フレームの特徴量を利用することに大きな効果はない．

さらに，チャネルのミスマッチがある C セットでは，クリーン音声状態の識別を用いるよりも，$\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$ の GMM を利用した方が性能が高い．これは，チャネルのミスマッチは特徴量の引き算によりキャンセルできるためだと考えられる．$\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$ の GMM の利用は，ヒューリスティックなやり方ではあるものの，チャネルのミスマッチが存在する可能性がある状況では，非常に実用的な方法であることが分かった．

HMM の学習条件として，clean 条件のものと multi 条件のものを比較すると，特に B セットにおいて，clean 条件の最も WER が低いもの（7.19）の方が，multi 条件の最も WER が低いもの（7.31）より，よい性能になっている．これは，提案手法では雑音抑圧の力が十分強いために，雑音環境にミスマッチがある場合には，NAT を行わない方が精度が高くなっていると解釈できる．

次に,「区分」に $\boldsymbol{v}_t(9)$ の GMM を用い,「線形変換」に $\boldsymbol{e}_t$ (9) の線形変換を用いる手法について，正則化パラメタを変えながら実験を行った結果を 表3.2 に示す．結果から，小さい正則化パラメタでは，正則化パラメタの違いによる精度の変化が小さいことが分かる．よって，正則化パラメタの設定に関しては，バリデーションセットを使って設定すれば，よい値が得られると予想される．なお，$\lambda = 0$ とすると，データ数が不足し，解析解の計算における逆行列が計算できなくなってしまう．

表 3.2: 正則化パラメタを変化させたときの AURORA2 データベースにおける WER (%) の平均.

			clean 条件				multi 条件			
区分	線形変換	λ	A	B	C	平均	A	B	C	平均
$\boldsymbol{v}_t(9)$	$\boldsymbol{e}_t$ (9)	1	7.84	8.31	10.67	8.59	6.93	8.57	9.12	8.02
$\boldsymbol{v}_t(9)$	$\boldsymbol{e}_t$ (9)	10^{-1}	6.92	7.55	9.55	7.70	6.28	7.89	8.37	7.34
$\boldsymbol{v}_t(9)$	$\boldsymbol{e}_t$ (9)	10^{-2}	6.51	7.26	**9.05**	7.32	6.10	7.53	8.25	7.10
$\boldsymbol{v}_t(9)$	$\boldsymbol{e}_t$ (9)	10^{-3}	**6.50**	**7.19**	9.42	**7.36**	**6.06**	7.37	**8.17**	**7.01**
$\boldsymbol{v}_t(9)$	$\boldsymbol{e}_t$ (9)	10^{-4}	6.74	7.32	10.57	7.74	6.10	**7.36**	8.39	7.06

3.5　まとめ

本章では，ステレオデータを学習データとして用いる区分的線形変換を用いた特徴量強調法として，クリーン音声状態識別に基づく区分の決め方と，ノイジー音声の特徴量と雑音特徴量の推定値を結合して前後数フレームの連結したものを特徴量にする方法，加えて特徴量の次元数が高くなった場合に線形変換の学習時に L2 正則化を導入する，DPLT という手法を提案した．AURORA2 データベースを用いた実験の結果，クリーン音声状態の識別，結合特徴量を線形変換に用いること，正則化にそれぞれ効果があり，SPLICE や NMN-SPLICE を越える精度が実現できることが分かった．

3.6　応用分野

本章では，クリーン音声状態の識別に基づき，前後数フレームの特徴量を入力とした区分的線形変換を用いた DPLT による特徴量強調法を提案したが，DPLT は，特徴量強調以外の分野にも応用することが可能である．

声質変換の分野では，A さんの声から B さんの声質の音声を作成するために，区分的線形変換の枠組みが広く利用されている．DPLT を声質変換に応用し，B さんの声状態の識別に基づいた前後数フレームの特徴量を入力とした区分的線形変換を用いることで，声質変換の性能を向上させられる（発表文献 [32] ）．

また電話音声のような狭帯域音声から，帯域を拡張した音声を作る研究でも，DPLT を利用することで，広帯域音声の状態の識別に基づいた前後数フレームの特徴量を入力とした区分的線形変換を用いることで，性能を向上させられる（発表文献 [33]）．

また，本論文では DPLT の区分的線形変換の入力特徴量として，ノイジー音声特徴量と，推定したノイズの特徴量を利用したが，それ以外の特徴量も利用することも可能である．例えば，他の雑音抑圧手法をかけた結果を入力として利用することができる（発表文献 [34]）．また，音声に関する特徴量だけでなく，例えば口唇画像特徴量を利用すること

で，マルチモーダル音声認識を行うことも可能である（発表文献 [11][43]）．

DPLT などの特徴量強調は，音声認識のフロントエンドのみならず，他の様々な音声アプリケーションのフロントエンドとしても利用できる．例えば，騒がしい教師で利用される外国語発音自動評価システムや（発表文献 [12]），騒がしいデモ開場で利用される声質変換システム（発表文献 [39]）などのフロントエンドとして利用できる．

第4章

ミスマッチがない場合にも頑健な特徴量強調

4.1　はじめに

第 3 章 では，雑音混入による音声認識精度の低下を防ぐため，雑音抑圧手法について述べてきた．これらにより，雑音が混入した音声の認識精度を向上させることができるが，雑音が混入していない条件では，逆に精度を低下させてしまう場合がある．そのため，雑音が混入していない場合に性能が劣化しないような特徴量強調の実現が望まれる．

そこで本節では，入力特徴量と音響モデルにミスマッチがあるかないかを表すコンディション変数を導入し，コンディションに依存させて特徴量強調を行うことで，ミスマッチがない条件にも頑健な手法を実現する．

4.2　提案手法

ある時刻フレームにおいて観測した特徴量を $\boldsymbol{y}$，それに対応するクリーン音声特徴量を $\boldsymbol{x}$ とおく．特徴量強調の目的は，$\boldsymbol{y}$ から，クリーン音声特徴量の推定値 $\hat{\boldsymbol{x}}$ を以下のように得ることである．

$$\hat{\boldsymbol{x}} = \operatorname*{argmax}_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) \tag{4.1}$$

提案手法では，ミスマッチがある場合 1，ない場合に 0 となるコンディション変数 c を導入し，(4.1) を以下のように展開する．

$$\hat{\boldsymbol{x}} = p(c=1|\boldsymbol{y}) \operatorname*{argmax}_{\boldsymbol{x}} p(\boldsymbol{x}|c=1,\boldsymbol{y}) + p(c=0|\boldsymbol{y}) \operatorname*{argmax}_{\boldsymbol{x}} p(\boldsymbol{x}|c=0,\boldsymbol{y}) \tag{4.2}$$

$$p(c=0|\boldsymbol{y}) = 1 - p(c=1|\boldsymbol{y}) \tag{4.3}$$

ここで，$\operatorname{argmax}_{\boldsymbol{x}} p(\boldsymbol{x}|c=0,\boldsymbol{y})$ は，ミスマッチがない場合なので，$\boldsymbol{y}$ とおくのが自然である．一方，$\operatorname{argmax}_{\boldsymbol{x}} p(\boldsymbol{x}|c=1,\boldsymbol{y})$ は，ミスマッチがある場合なので，通常の特徴量強調結果を代入すればよい．特徴量強調手法には任意の手法を採用してよく，本稿では前節までで紹介したステレオベース特徴量強調を利用する．

$p(c=1|\boldsymbol{y})$ の計算のために，生成モデルを置く．クリーン音声特徴量を θ_0 をパラメタに持つ生成モデル，雑音を重畳した音声特徴量を θ_1 をパラメタに持つ生成モデルでモデル化する．本論文では，具体的に，特徴量が時間によらず GMM に従って生成されるとモデル化する．θ_0 と θ_1 は，予め学習データを用いて学習しておく．

T フレーム分の一発話 $\boldsymbol{y}_{1:T}$ が与えられたとき，その発話のコンディション変数が $c=1$ となる事後確率は，事前確率を $p(c=0)=p(c=1)=0.5$ と仮定すると，標準シグモイド

関数 $\sigma(x) = (1 + \exp(-x))^{-1}$ を用いて以下のように計算できる．

$$
\begin{aligned}
p(c=1|\boldsymbol{y}_{1:T}) &= \frac{p(\boldsymbol{y}_{1:T}|c=1)p(c=1)}{p(\boldsymbol{y}_{1:T}|c=1)p(c=1) + p(\boldsymbol{y}_{1:T}|c=0)p(c=0)} && (4.4)\\
&= \frac{\prod_{t=1}^{T} p(\boldsymbol{y}_t;\theta_1)}{\prod_{t=1}^{T} p(\boldsymbol{y}_t;\theta_1) + \prod_{t=1}^{T} p(\boldsymbol{y}_t;\theta_0)} && (4.5)\\
&= \frac{1}{1 + \prod_{t=1}^{T} \frac{p(\boldsymbol{y}_t;\theta_0)}{p(\boldsymbol{y}_t;\theta_1)}} && (4.6)\\
&= \sigma\left(\sum_{t=1}^{T} (\ln p(\boldsymbol{y}_t;\theta_1) - \ln p(\boldsymbol{y}_t;\theta_0))\right) && (4.7)
\end{aligned}
$$

ただし，$\boldsymbol{y}_t$ は時刻 t に観測した特徴量，$\ln p(\boldsymbol{y}_t;\theta_0)$ および $\ln p(\boldsymbol{y}_t;\theta_1)$ は，それぞれ θ_0，θ_1 に対応する GMM の $\boldsymbol{y}_t$ の対数尤度である．

本論文ではさらに，$p(c=1|\boldsymbol{y}_{1:T})$ が 0.5 より大きいか小さいか（シグモイド関数の中身の正負）によって，発話単位でコンディション変数を 1 か 0 に一意に定める．すなわち，以下のように特徴量強調を行う．

$$
\hat{\boldsymbol{x}} = \begin{cases} \boldsymbol{y} & \text{if } \sum_{t=1}^{T} (\ln p(\boldsymbol{y}_t;\theta_1) - \ln p(\boldsymbol{y}_t;\theta_0)) \leq 0 \\ \operatorname{argmax}_{\boldsymbol{x}} p(\boldsymbol{x}|c=1,\boldsymbol{y}) & \text{if } \sum_{t=1}^{T} (\ln p(\boldsymbol{y}_t;\theta_1) - \ln p(\boldsymbol{y}_t;\theta_0)) > 0 \end{cases} \tag{4.8}
$$

このように一意にコンディションを決定することから，本論文ではこの手法を，CN 識別（Clean 音声と Noise を含む音声の識別）と呼ぶ．

CN 識別のメリットに，バックエンドを切り替えられる点がある．一般的に，クリーン音声の認識はクリーン音声で学習した音響モデルが最も精度が高く，雑音重畳音声の認識は雑音重畳音声に特徴量強調をかけた音声を用いて Noise Adaptive Training（NAT）した音響モデルの精度が高くなる．そこで，発話単位のコンディションが 0 の場合にはクリーン音声で学習した音響モデルを用い，発話単位のコンディションが 1 の場合には NAT で学習した音響モデルを用いることにする．これにより，すべての場合で高い精度が得られると考えられる．

4.3　実験

実験に用いるデータベースには，第 3 章 と同じ AURORA2 データベースを利用する [61]．

コンディション変数の導入によるミスマッチがない場合にも頑健な手法の評価を行うため，以下の三つの条件で実験を行った．clean 条件および multi 条件がコンディション変数を使わない手法，CN 識別が提案手法である．

clean 条件　クリーン音声のみを用いて単語単位の HMM/GMM を最尤推定する．特徴量

表 4.1: AURORA 2 における WER の平均（%）．クリーンは clean1-4 の平均，A，B，C はそれぞれのテストセットにおける SNR 0-20 の平均を表す．

	クリーン	A	B	C
clean 条件	**0.33**	23.26	20.97	19.55
multi 条件	0.57	**5.66**	**6.45**	**6.25**
CN 識別	**0.33**	**5.66**	6.46	**6.25**

には PLP + Δ + ΔΔ を fMLLR で話者適応し CMN したものを利用する．

multi 条件　SNR 5,10,15,20,∞ [dB] で雑音が付加された音声を用いて単語単位 HMM/GMM を NAT する．特徴量には，PLP + Δ + ΔΔ を [64] に倣って fMLLR で話者適応し，本論文で提案したステレオベース特徴量強調（以降，Discriminative Piecewise Linear Transformatin; DPLT と呼ぶ）を行い，最後に HEQ [26] で特徴量正規化をしたものを利用する．

CN 識別　コンディション変数の推定結果に従い，clean 条件と multi 条件を選択して認識を行う．

この実験条件は，第 3 章 の実験と異なり，DPLT だけでなく PLP，fMLLR，HEQ などの要素技術を導入しているため，ベースラインの精度が向上することになる．また実用性をより強く意識し，雑音を推定するための計算量を小さくするために，音声ファイルの先頭 10 フレームの平均値を，全フレームにおける推定値 $\hat{\boldsymbol{n}}$ として利用した．特徴量抽出の窓には，広く用いられているフレーム長 25 msec，シフト長 10 msec とした．

DPLT の学習データには，クリーン音声を含む HMM/GMM の学習データ 8440 発声をステレオデータにしたものを用いた．クリーン音声状態識別のためにソフトな LDA をかけた特徴量の GMM を利用し，LDA の入力特徴量 $\boldsymbol{d}_t$ としては，前後フレームを用いず，当該フレームの雑音重畳音声特徴量 $\boldsymbol{y}_t$ と雑音推定値 $\hat{\boldsymbol{n}}$ の結合ベクトルを用いた．線形変換の特徴量としては，雑音重畳音声の当該フレームを $\boldsymbol{y}_t$ として，$\boldsymbol{y}_{t-5}, \boldsymbol{y}_{t-4}, \boldsymbol{y}_{t-3}$ の平均，$\boldsymbol{y}_{t-2}, \boldsymbol{y}_{t-1}$ の平均，$\boldsymbol{y}_t$，$\boldsymbol{y}_{t+1}, \boldsymbol{y}_{t+2}$ の平均，$\boldsymbol{y}_{t+3}, \boldsymbol{y}_{t+4}, \boldsymbol{y}_{t+5}$ の平均，さらに雑音の推定値 $\boldsymbol{n}$ とバイアス項 1 をすべて結合したものを利用した．正則化パラメタは 10^{-3} とした．

fMLLR による話者適応は，一名につき約 10 発声分のデータを，教師なしデータとして利用して行った．音響モデルの学習の際にも同様の変換を行なっており，音響モデルの HMM/GMM は SAT した．fMLLR のモデルとしては，HMM/GMM でなく，別にクリーン音声のみで学習した 1024 混合の GMM を利用した [64].

コンディション変数の推定に用いる生成モデルには，32 混合の GMM を用い，特徴量には PLP + Δ + ΔΔ を CMN したものを利用した．θ_1 の学習には SNR 5,10,15,20 [dB] で雑音が付加された 6752 発声を，θ_0 の学習にはそれと同じ音声データで雑音が含まれていない 6752 発声を利用した．

実験結果を 表 4.1 に示す[1]．clean 条件はクリーン音声を認識する上では性能が高いが，雑音のミスマッチがあると精度が低くなってしまう．multi 条件は A，B，C セットを認識する上では性能が高いが，雑音のミスマッチがない条件では精度が低くなってしまう．clean 条件でも multi 条件でも，雑音推定値を簡略化して計算量を小さくしたにも関わらず， fMLLR による話者適応や HEQ の効果により，前節と比較しても高い認識率が得られている．提案手法である CN 識別を用いると，ミスマッチのあるなしに関わらず，ほぼ最高性能を実現することができる．雑音環境オープンテストである B セットでも，multi 条件と CN 識別の差はほぼないため，コンディション変数の推定は雑音環境に依らず適切に行えていることが分かる．

4.4　過去の文献との比較

表 4.2: 様々な文献における AURORA2 データベースにおける WER (%) の平均．

	clean 条件				multi 条件			
手法	A	B	C	平均	A	B	C	平均
AFE [14]	12.56	13.00	14.45	10.68	7.88	8.04	9.43	8.26
Extended VTS 適応 [13]	7.0	7.2	6.9	7.1	-	-	-	-
VTS-NAT [28]	7.21	6.74	7.41	7.06	6.34	6.23	6.11	6.25
Exemplar-based method [16]	4.1	5.6	-	-	3.1	5.0	-	-

ここまで，雑音抑圧手法として DPLT を，またミスマッチがある場合にも頑健な手法として CN 識別を提案してきた．ここで，過去に提案された様々な雑音に頑健な音声認識手法について，文献に載せられている結果を調査し，提案手法と比較する．当然実験条件の細かな違いはあるが，おおよその傾向はつかむことができる．

調査の結果を 表 4.2 に示す．ただし，論文に実験結果が示されていない場合は - で表した．まず，AFE と比べると，今回の提案手法の方が十分に精度が高いことがわかるまた HMM/GMM の雑音適応の一つの state-of-the-art である [13] や [28] と比べても，提案手法は同等以上の精度を実現できていることが分かる．特に，雑音環境クローズド条件（A セット）では，VTS を用いるよりも，提案手法の方が精度が高い．また，現在最も AURORA2 データベースで WER が低いのは [16] の手法であり，これと比べると提案手法は精度で劣っている．しかしながら [16] の exemplar-based method は，HMM/GMM の雑音適応以上に計算コストが非常に大きく，実時間で動作させるのはほぼ不可能である．以上の調査から，提案手法は，現実的な計算時間で実行できる雑音ロバスト音声認識手法の，一つの state-of-the-art であると言うことができる．

[1]ただしここではそれぞれのセットの WER の平均値のみを載せている．詳細な WER は付録にある．

4.5　まとめ

本章では，ミスマッチがない条件で性能が低下しない特徴量強調を実現するために，発話単位でコンディション変数を推定し，バックエンドの音響モデルを切り替えながら音声認識を行う手法を提案した．この手法を DPLT と併用した結果，ミスマッチのあるなしに関わらず，高い認識精度が得られることが分かった．

第5章

識別的リランキングにおける構造的表象の利用

5.1　はじめに

音声認識をはじめ，構文解析，機械翻訳など様々自然言語処理タスクにおいて，識別的リランキング（discriminative reranking）が，さらなる精度向上を実現する手法として注目されている [65]．識別的リランキングとは，何らかの手法で複数の解候補を選んだ後，識別モデルを使って解候補をリランキングする手法である．

識別的リランキングの一つの利点は，任意の特徴量を利用できることである．例えば音声認識の音響モデルに HMM/GMM や HMM/DNN を利用する場合，特徴量としては時間インデックス t ごとの局所的特徴量しか利用できない．しかし識別的リランキングでは，時間的に局所的特徴量に限らず，任意の長時間単位にまたがる特徴も利用できる．

音声認識の識別的リランキングに関しては，従来，単語 n-gram カウントやニューラルネットワークを用いた言語モデルの確率など，言語的な特徴量を利用した「識別的言語モデル」が広く用いられている．また，音響的な特徴量を利用する手法としては，SCARF が有名で，例えばデュレーションなど，単語単位にまたがる音響的な特徴量などが利用されている．識別的言語モデルと SCARF などの識別モデルは，目的や手法が似ているにもかかわらず，これまで比較的独立して発展してきている[1]．

本研究ではこの状況を鑑みて，識別的言語モデルで広く用いられているモデルを利用し，長時間にわたって定義される音響特徴量の利用する手法を提案する．これまでの SCARF などの音響的特徴量を利用した識別モデルでは，識別モデルとして segmental CRF のような構造的なモデルを利用することで，精度を高めようと発展してきた．そのため，特徴量として単語単位のものは使われても，文にまたがる特徴量は利用されてこなかった．一方，識別的言語モデルでは，識別モデルとしては文全体によって定義される特徴量を利用した識別モデルが利用され，学習データの作成や，feature engineering （特徴量設計）に関する技術が発展してきた．本研究では，識別的言語モデルのやり方にそって，識別モデルとしては文単位の線形識別モデルを用い，特徴量として長時間音響特徴量を導入することで，さらなる精度向上を狙うものである．

本研究では，具体的な長時間音響特徴量として，音声の構造的表象を利用することを提案する．構造的表象は，一発声に含まれる音素と音素の距離から構成されるもので，話者の違いなどの非言語的特徴のミスマッチに頑健な特徴量である．しかし文などの長時間にわたって定義される特徴量であるために，大語彙音声認識のデコーディングアルゴリズムと相性が悪く，これまで大語彙音声認識に用いられた例は存在していなかった．一方，今回の識別的リランキングの枠組みを用いれば，デコーディングは前段階で終了しているため，構造的表象のような長時間音響特徴量も利用することが可能となる．

[1]近年になって，それらを統一的に扱ったサーベイ論文が登場するなどしている [38]．

5.2　大語彙音声認識の識別的リランキング

大語彙音声認識の識別的リランキングのモデル化には数多くのやり方があるが，本稿では，N 個の認識結果の仮説を入力とし，線形識別モデルを用いて N ベストリランキングを行うタイプのものを用いる．

まず，線形識別モデルに用いる特徴量を定義する．具体的には，観測した音声特徴量系列 $\boldsymbol{X}$，その n 番目の単語系列仮説 W_n から，任意の特徴量ベクトル $\boldsymbol{\Phi}(\boldsymbol{X}, W_n)$ を定義する．例えば識別的言語モデルの枠組みでは，

$$\boldsymbol{\Phi}(\boldsymbol{X}, W_n) = \begin{bmatrix} \text{“a” という単語が } W_n \text{に含まれる数} \\ \text{“the” という単語が } W_n \text{に含まれる数} \\ \vdots \end{bmatrix}. \tag{5.1}$$

のように，仮説に含まれる単語のカウントを特徴量として利用できる．また他にも，例えば PLP の時間平均など $\boldsymbol{X}$ に関する特徴量や，$\boldsymbol{X}$ と W_n の両方に依存する特徴量など，任意の特徴量が利用できる．

次に，上記で定義した任意の特徴量を用い，以下の線形識別モデルで最終的な認識結果 W^* を選択する．

$$W^* = \underset{W_n \in \mathrm{NBEST}(\boldsymbol{X})}{\operatorname{argmax}} \boldsymbol{\alpha} \cdot \boldsymbol{\Phi}(\boldsymbol{X}, W_n) + \phi_0(\boldsymbol{X}, W_n) \tag{5.2}$$

ここで，NBEST($\boldsymbol{X}$) は，$\boldsymbol{X}$ から HMM/GMM などを利用して音声認識した結果得られる N 個の単語系列仮説の集合を表す．$\phi_0(\boldsymbol{X}, W_n)$ は，N ベストを出力した際に得られる対数尤度スコアである[2]．$\boldsymbol{\alpha}$ は線形識別モデルのパラメタであり，次元数は $\boldsymbol{\Phi}(\boldsymbol{X}, W_n)$ と同じで，各特徴に対する重みとなっている．$\boldsymbol{\alpha} = \mathbf{0}$ であれば，$\phi_0(\boldsymbol{X}, W_n)$ がそのままスコアになるため，N ベストリランキングを行わずに 1 位仮説を出力することに相当する．

$\boldsymbol{\alpha}$ は，例えば 図5.1 に示す平均化パーセプトロンの一種を用いて学習する．$\overline{W_i}$ と $\underline{W_i}$ は，観測音声 $\boldsymbol{X}_i$ に対する N 個の仮説の中で，最も WER が高かった仮説，低かった仮説をそれぞれ表す．このアルゴリズムの中心的なアイディアは，WER が最も高い仮説にペナルティを与え，逆に WER が最も低い仮説に報酬を与える，というものである [66]．アルゴリズムの最後に $\boldsymbol{\alpha} = \Sigma_{i,t}\boldsymbol{\alpha}_i^t/IT$ と平均をとってるが，これはマージンを大きくし汎化性能を向上させる効果がある [67]．

[2]対数の足し算を行っていることから，この線形モデルは対数線形モデル（ログリニアモデル）と呼ばれることも多い．

Input: Training samples $(\boldsymbol{X}_i, \overline{W_i}, \underline{W_i})$ for $i = 1 \ldots I$
Initialization: $\boldsymbol{\alpha}_0^I = \mathbf{0}$
1: **for** $t = 1 \ldots T$ **do**
2: $\quad \boldsymbol{\alpha}_t^0 = \boldsymbol{\alpha}_{t-1}^I$
3: $\quad$ **for** $i = 1 \ldots I$ **do**
4: $\quad\quad$ **if** $\boldsymbol{\alpha}_t^{i-1} \cdot \boldsymbol{\Phi}(\boldsymbol{X}_i, \overline{W_i}) + \phi_0(\boldsymbol{X}_i, \overline{W_i}) > \boldsymbol{\alpha}_t^{i-1} \cdot \boldsymbol{\Phi}(\boldsymbol{X}_i, \underline{W_i}) + \phi_0(\boldsymbol{X}_i, \underline{W_i})$ **then**
5: $\quad\quad\quad \boldsymbol{\alpha}_t^i = \boldsymbol{\alpha}_t^{i-1} + \lambda \left(\boldsymbol{\Phi}(\boldsymbol{X}_i, \underline{W_i}) - \boldsymbol{\Phi}(\boldsymbol{X}_i, \overline{W_i}) \right)$
Output: $\boldsymbol{\alpha} = \Sigma_{i,t} \boldsymbol{\alpha}_i^t / IT$

図 5.1: 平均化パーセプトロンアルゴリズムの一種．上線，下線が引かれた W_i は，i 番目の音声の N ベストリストの中で，それぞれ最も高い（悪い） WER と，低い（良い）WER のものを表す．I は学習データの数である．T はアルゴリズムの繰り返し回数である．λ は学習率で，本研究では予め一つの値に固定する．

5.3　音声の構造的表象

5.3.1　f-divergence

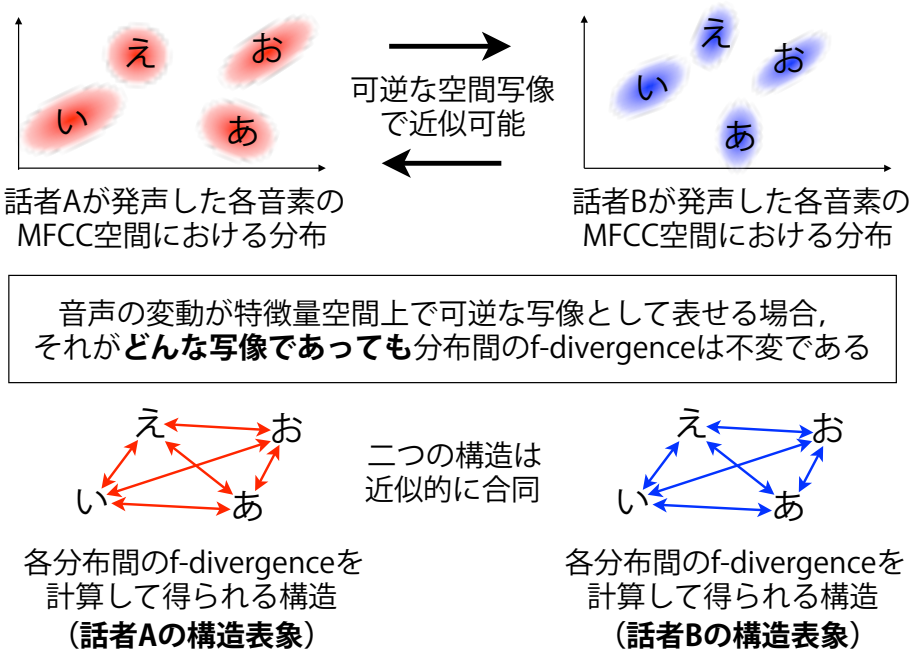

図 5.2: 音声の構造的表象

音声の構造的表象の概念図を図 5.2 に示す．ある二つの分布に任意の一対一対応変換を施しても，その分布間距離尺度の一種である f-divergence は不変であることが証明されてい

る [68]．音声の構造的表象とは，ある話者の発声を音素などの単位で特徴量空間上で分布化し，それらすべての分布間の f-divergence を計算することで得られる構造のことである．

ここで，話者の違いは，特徴量空間の一対一対応変換で近似することができる．例えば MLLR 適応では，これを決定木で区分的にした線形変換と仮定している．構造的表象は，このような変換に不変であるため，話者の違いに高い頑健性を持つ．話者の違いの他にも，例えばマイクの違いや伝送特性の違いなども，特徴量の線形変換として近似できるため，構造的表象は高い頑健性を持つ．

5.3.2　音声の構造的表象を用いた単語音声認識

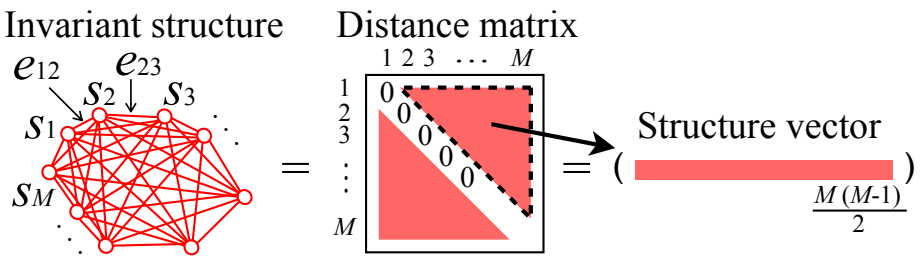

図 5.3: 音声の構造的表象のベクトル表現

構造的表象に関する記号を 図 5.3 を用いて定義する．構造的表象は，M 個のノードからなる．それぞれのノードを，$\{s_i\}_{i=1\cdots M}$ で表す．それぞれのノードは，音響イベントの分布であり，具体的には音素 HMM の各状態などが対応する．次に構造のエッジ長を，$\{e_{ij}\}_{1\leq i\leq M, i<j\leq M}$ で表す．エッジ長は，二つのノード間の f-divergence である．音声の構造的表象は，数学的には距離行列として表現できる．もし，f-divergence として対称な距離尺度を選べば，距離行列の上三角成分だけで情報をすべて表現できることになる．このような f-divergence として，Bhtaccharyya Distance の平方根 $\sqrt{\mathrm{BD}}$ がよく用いられる．距離行列の上三角成分をベクトルに並べ直したものを，構造ベクトルと呼ぶ．構造ベクトルは，$\frac{M(M-1)}{2}$ 次元のベクトルである．

音声の構造的表象を用いた孤立単語音声認識の枠組みを 図 5.4 に示す．図 5.4 の左側は，入力の孤立単語発声から構造ベクトルを抽出する方法を示している．まず，入力された孤立単語発声から MFCC などの短時間特徴量系列を抽出する．次に，得られた一つの短時間特徴量系列のみから，left-to-right HMM を学習する．この際，HMM の状態数は M 個に予め固定されており，各状態の出力する分布は正規分布とする．そして，学習された HMM 各状態の持つ正規分布を，構造的表象のノードとみなして利用する（HMM の状態遷移確率は捨てる）．ここで，少ないデータから分布を推定したことによる不安定性を取り除く

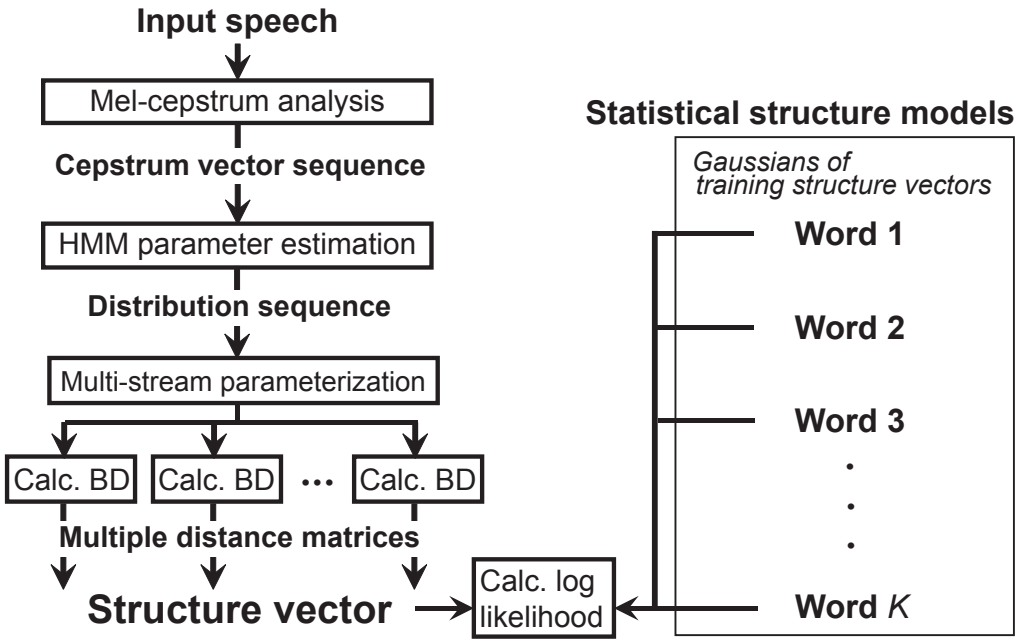

図 5.4: 音声の構造的表象を用いた孤立単語音声認識の枠組み

ため，大量に集めた不特定話者・不特定音素の正規分布との内挿値をとる．次に正規分布系列を，特徴量の次元方向に分割し，マルチストリーム化する．この処理は，構造的表象の強すぎる不変性に制約条件を加え，音声認識に必要な情報を増やす効果がある．これ関しては，[69] を参照されたい．その後，各ストリームごとに，f-divergence を計算することで構造的表象を抽出し，それを構造ベクトルにする．ここで，マルチストリーム化をしているため，構造ベクトルの次元数はストリーム数倍となる．以上で入力の孤立単語発声から構造ベクトルを抽出することができたので，これを単語ごとの構造ベクトルの生成モデルと比較し，最も対数尤度が大きいモデルに対応する単語を音声認識結果として出力する．ここで当然ながら，生成モデルではなく識別モデルを利用することも可能である．実際，特徴量を LDA により識別的に次元圧縮することで，精度が向上できることが示されている [59].

上記の手法を用いれば，音声の構造的表象を単語音声認識に利用することできるが，その精度は，適応などを全く行なっていない HMM/GMM と比べても低い．f-divergence の不変性に関しても，分布の形状を正規分布と仮定したり，分布の推定精度が不十分であったりすることが原因で，話者の違いに対する不変性は完全には成り立たない．また，大語彙音声認識には利用できない問題もある．

5.3.3　音声の構造的表象を用いた外国語発音評価

音声の構造的表象を用いるメリットは，話者が発声した二つの音素の，違いに関する情報を明示的に利用できることである．例えば，日本人が英語を発声するときに，/l/ と /r/

を区別せずに発声してしまう話者が多い．このとき，音声の構造的表象のように，音素間の違いを明示的に利用していれば，このような発音誤りを検出するのに有効であると考えられる．

実際，外国語発音評価タスクにおいては，音声の構造的表象を用いることで，HMM/GMM を用いる場合を越える精度を実現している [60]．また，HMM/GMM で得られたスコアと，音声の構造的表象で得られたスコアを組み合わせることで，さらなる精度向上が実現できることも示されている．

5.4　提案手法

音声認識の識別的リランキングの一つの利点は，長時間にわたって定義される任意の特徴量を利用できる点にある．しかしこれまで，音響特徴量の長時間にわたって定義される特徴量はほとんど検討がなされていない．このような特徴量を用いることで，従来利用されていなかった音声認識の情報が利用でき，精度を向上させることができると考えられる．

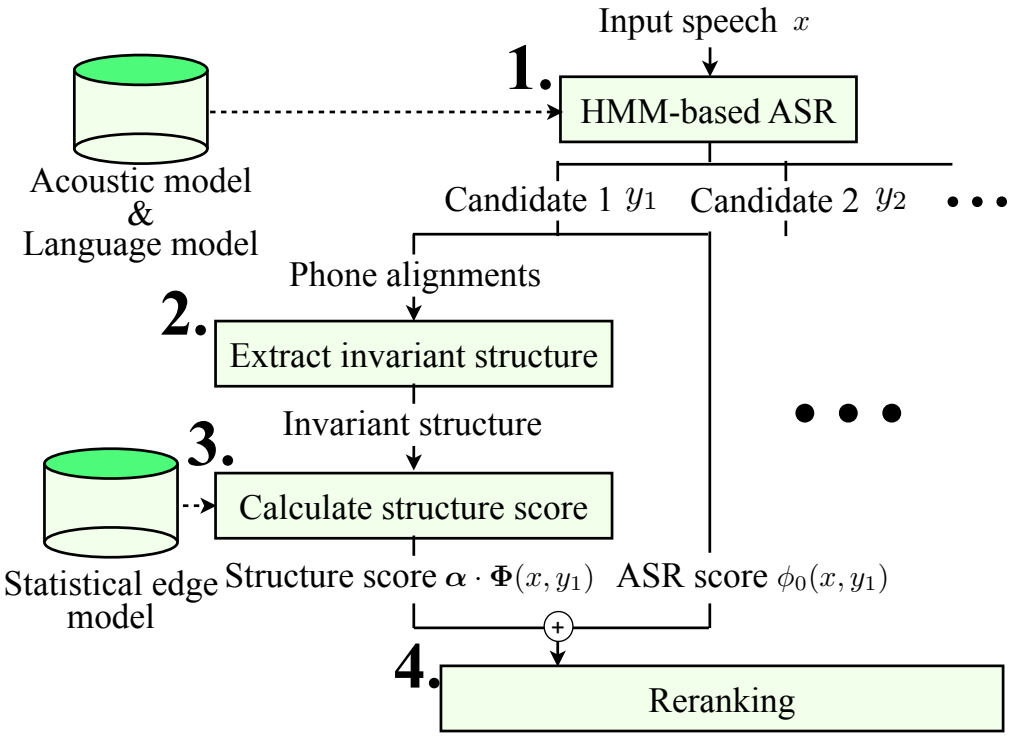

図 5.5: 音声認識の識別的リランキングに構造的表象を用いる手法の概略

本研究では，音声認識の識別的リランキングの特徴量として，構造的表象を利用することを提案する．提案手法の概略を 図 5.5 に示す．図 5.5 の 1〜4 の番号は，以下のサブセクションに対応している．

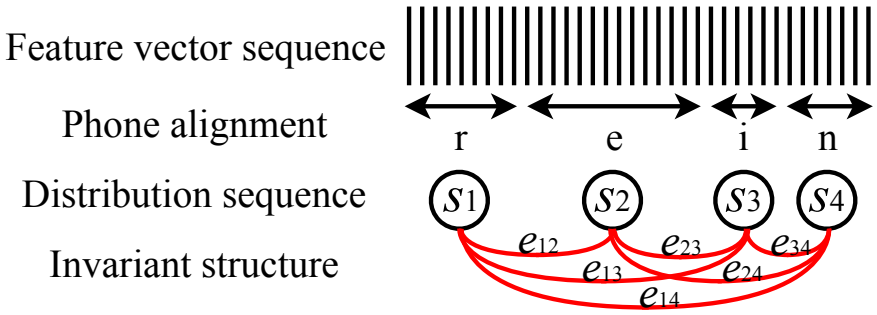

図 5.6: 音声の構造的表象を強制アライメント結果から抽出する方法

5.4.1　HMM ベースの音声認識

まず，広く利用されている HMM ベースの音声認識システムを用い，対数尤度の高い上位 N 個の仮説を得る．ここでそれぞれの仮説ごとに，対数尤度スコア $\{\phi_0(\boldsymbol{X}, W_n)\}_{n=1\cdots N}$ を出しておく．また，W_n を用いて，強制アライメントを行い，音素アライメント結果も得ておく．

5.4.2　音声の構造的表象の抽出

次に，N 個の仮説それぞれから構造的表象を抽出する．図 5.6 に，仮説 [r e i n] から構造的表象を抽出する方法を示す．まず，音素アライメントを利用して，音素ごとに正規分布を推定する．次にそれらの正規分布間の f-divergences$\{e_{ij}\}_{1<i<j<M}$ を計算して，構造的表象を抽出する．

正規分布の平均と分散を推定する際，非常に少ないデータからそれらを計算しなければならないため，最尤推定でパラメタを推定すると，特に分散の値が不安定になってしまうことが多い．そこで，分散に関しては，学習データを用いて音素毎に分散を計算しておき，それを常に利用する．平均に関しては，最尤推定で決定する．

5.4.3　構造スコアの計算

次に，構造的表象を構成する各エッジの尤もらしさのスコアを特徴量として，識別的リランキングを行う．

まず「各エッジの尤もらしさ」を計算するために，音素ペアごとに統計的エッジモデル (Statistical Edge Model; SEM) を学習しておく．図 5.7 の左側に，SEM の学習プロセスを示す．SEM の学習では，二つの音素ペアをラベルとして，1 次元の f-divergence の長さの生成モデルを GMM で学習する．音素が P 種類あった場合，SEM は $P(P-1)/2$ 個学習することになる．

図 5.7 の右側に，単語系列仮説と構造的表象 ($\{e_{ij}\}_{1<i<j<M}$) が入力されたときに，構造スコアを計算するプロセスを示す．まず，構造的表象のそれぞれのエッジ e_{ij} に対して，以

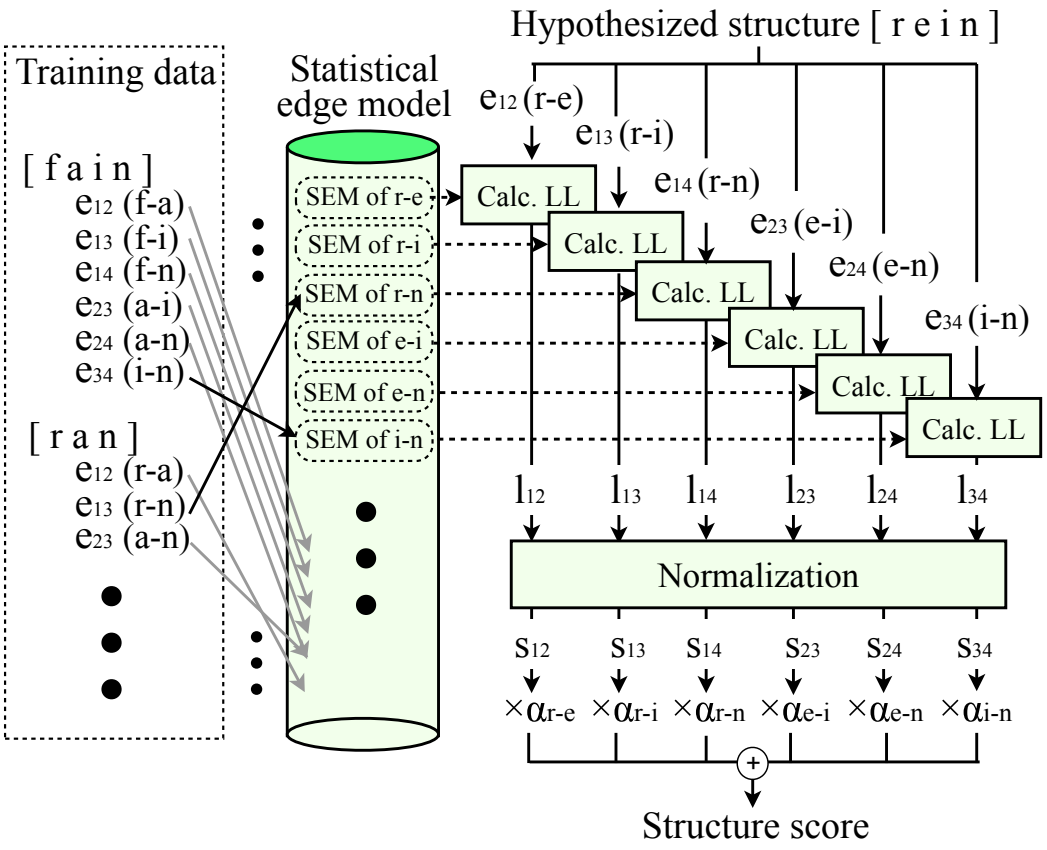

図 5.7: 統計的エッジモデル（SEM）の作成法．LL は，対数尤度（Log likelihood）の略である．

下のように SEM の対数尤度を計算する．

$$l_{ij} = \log \sum_{k=1}^{K} w_{p_{ij}}^{k} \mathcal{N}(e_{ij}; \mu_{p_{ij}}^{k}, \sigma_{p_{ij}}^{k}) \tag{5.3}$$

ここで p_{ij} は，e_{ij} に対応する音素ペア ID（Phoneme-Pair ID; PPID）を表す．$w_{p_{ij}}^{k}$, $\mu_{p_{ij}}^{k}$, $\sigma_{p_{ij}}^{k}$ は，PPID が p_{ij} の SEM（K 混合の GMM）の k 番目のコンポーネントの重み，平均，分散共分散行列である．そして得られた $\{l_{ij}\}_{1<i<j<M}$ を，duration，音素数 M で以下のように正規化する．

$$s_{ij} = \frac{f_i + f_j}{M - 1} l_{ij} \tag{5.4}$$

ここで f_i, f_j は i 番目，j 番目の音素にアライメントされたフレーム数で，音素アライメント結果から簡単に計算することが出来る．

最後に，最終的なスコアを $\boldsymbol{\alpha} \cdot \boldsymbol{\Phi}(\boldsymbol{X}, W_n)$ で計算する．ここで，$\boldsymbol{\alpha}$ と $\boldsymbol{\Phi}(\boldsymbol{X}, W_n)$ は，常に $\frac{P(P-1)}{2}$ 次元となるベクトルであり，それぞれの要素は PPID に対応する．$\boldsymbol{\Phi}(\boldsymbol{X}, W_n)$

は，それぞれの PPID に対応する正規化エッジスコア $\{s_{ij}\}_{1<i<j<M}$ の和で，以下のように定義される．

$$\boldsymbol{\Phi}(x,y) = \begin{bmatrix} \sum_{i,j} s_{ij} \text{ if } p_{ij} = 1 \text{ , otherwise } 0 \\ \sum_{i,j} s_{ij} \text{ if } p_{ij} = 2 \text{ , otherwise } 0 \\ \vdots \\ \sum_{i,j} s_{ij} \text{ if } p_{ij} = \frac{P(P-1)}{2} \text{ , otherwise } 0 \end{bmatrix} \tag{5.5}$$

もし特定の PPID が仮説 W_n の構造的表象に含まれなければ，その PPID に対する特徴は 0 になるため，$\boldsymbol{\Phi}(\boldsymbol{X}, W_n)$ はスパースなベクトルとなる．また $\boldsymbol{\alpha}$ は，先に説明した 図5.1 のアルゴリズムで予め学習しておく．そのため $\boldsymbol{\alpha}$ は，エラーレートを削減するためにどの音素ペア（PPID）のスコアを重要視すればよいかの度合い，と解釈できる．

5.4.4　リランキング

最後に，HMM ベースの音声認識から出力されたスコア $\phi_0(\boldsymbol{X}, W_n)$ と，構造的表象から得た構造スコアを組み合わせ，式 (5.2) を用いてリランキングを行う．

5.5　実験

5.5.1　実験条件

提案手法の有効性を検証するために，日本語の連続数字音声認識，大語彙音声認識の二つの実験を行った．実験条件を 表5.1, 表5.2 にまとめる．なお 10 ベストオラクルとは，10 ベストリストの中で最も誤り率が低いものを選択できた場合の結果で，識別的リランキングの枠組みにおける上限値である．また大語彙音声認識実験においては，WER のかわりに文字誤り率（Character Error Rate; CER）を用いた．CER を用いる理由は，日本語は単語分割に曖昧性があるためである．そのため $\boldsymbol{\alpha}$ の学習時の $\overline{W_n}$，$\underline{W_n}$ には，それぞれ，10 ベストの中で WER・CER が最も高い・低いものを利用することになる．

音響モデルにひあ HMM/GMM を用い，10-ベストの仮説と，音素アライメントを出力した．HMM/GMM は音素単位で学習され，決定木を用いて状態クラスタリングを行っている．構造的表象を得るための分布の特徴量には，13 次元の PLP 特徴量を使い，3 状態 left-to-right HMM の 2 状態目に対応する部分のみを用いてガウス分布の平均を ML 推定した．ガウス分布の分散に関しては，少ないデータから分布を学習するため，音素ごとに分散を予め設定しておき，それを常に利用している．f-divergence としては，バタチャリヤ距離の平方根 $\sqrt{\mathrm{BD}}$ を用いた．SEM としては，16 混合の GMM を用いた．

表 5.1: 日本語の連続数字音声認識の実験条件

発声内容	1 から 11 回の日本語連続数字読み上げ
HMM 学習データ	27.5 時間 / 667 話者 / 17316 発声
SEM 学習データ	27.5 時間 / 667 話者 / 17316 発声
$\boldsymbol{\alpha}$ 学習データ	5.0 時間 / 520 話者 / 3977 発声
テストデータ	1.5 時間 / 100 話者 / 7382 発声
HMM 状態数	500
HMM ガウシアン数	15000
monophone 数 (P)	18
monophone-pair 数	136
言語モデル	0 から 9 の 10 の数字，終端記号を 当確率で出力するユニグラム
ベースライン WER	1.09% (S=67, I=140, D=14 / 20303)*
10 ベストオラクル	0.75% (S=59, I=85, D=9 / 20303)*

* S: 置換誤り数, I: 挿入誤り数, D: 削除誤り数

表 5.2: 日本語の大語彙音声認識の実験条件

発声内容	日本語読み上げ音声
HMM 学習データ	352 時間 / 1325 話者 / 196475 発声
SEM 学習データ	24 時間 / 100 話者 / 13112 発声
$\boldsymbol{\alpha}$ 学習データ	30 時間 / 164 話者 / 16733 発声
テストデータ	1.5 時間 / 20 s 話者 / 600 発声
HMM 状態数	5000
HMM ガウシアン数	150000
monophone 数 (P)	57
monophone-pair 数	1596
言語モデル	Modified Kneser-Ney smoothing を用いた単語 2-gram
単語数	104262
ベースライン CER	3.59% (S=422, I=56, D=64 / 15096)*
10 ベストオラクル	1.32% (S=161, I=15, D=24 / 15096)*

* S: 置換誤り数, I: 挿入誤り数, D: 削除誤り数

5.5.2　結果

図 5.8 と 図 5.9 に，連続数字音声認識の WER 大語彙連続音声認識の CER をそれぞれ示す．図の横軸は $\boldsymbol{\alpha}$ 学習の繰り返し回数 T を表す．$\boldsymbol{\alpha}$ の初期値を $\mathbf{0}$ としているため，$T = 0$ は，ベースラインシステムのエラーレートを表す．λ は学習率であり，$0.1, 0.2, 0.5$ の三種類で実験を行った．

提案手法は，どのような λ，T においても，ベースラインの精度を上回る精度となった．

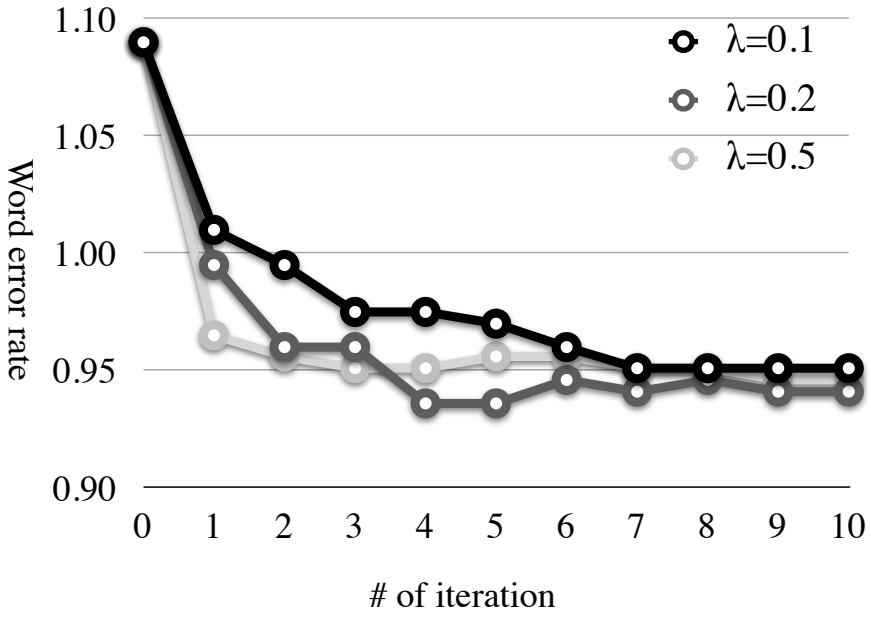

図 5.8: 日本語の連続数字音声認識の WER

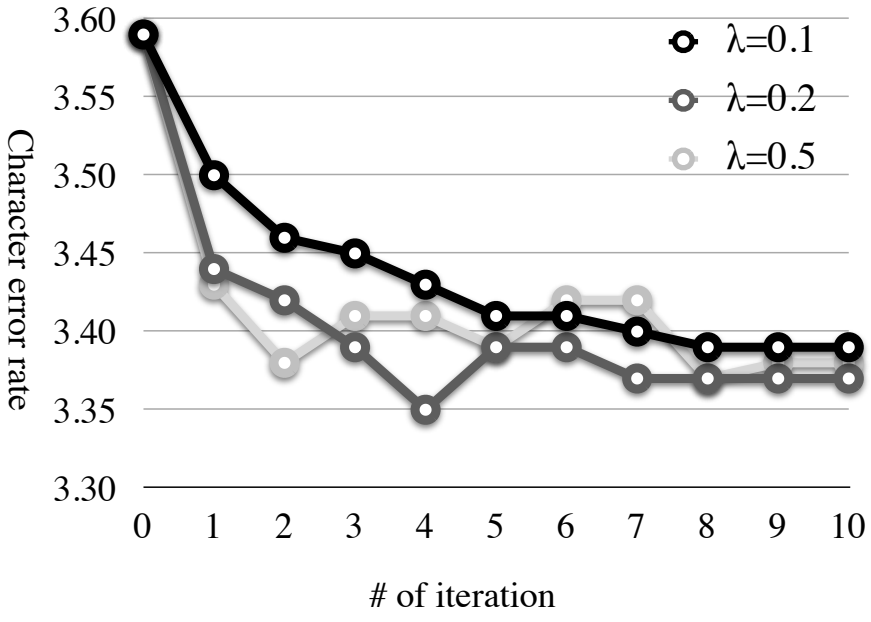

図 5.9: 日本語の大語彙音声認識の CER

連続数字読み上げ音声の実験に関しては，$\lambda = 0.2$，$T = 4$ または 5 のとき，WER は 0.94% と最小になり，ベースラインシステムから 14.1% の WER 削減を達成した．このとき，置換誤りは 67 から 64 に，挿入誤りは 140 から 113 に，削除誤り 14 から 13 に，それぞれ減

少した．大語彙音声認識に関しては，$\lambda = 0.2$，$T = 4$ のとき，CER は 3.35% と最小になりベースラインシステムから 6.69% の CER 削減を達成した．このとき，置換誤りは 422 から 401 に，挿入誤りは 56 から 44 に，削除誤り 64 から 61 に，それぞれ減少した．

5.6　まとめ

本論文では，大語彙音声認識の識別的リランキングにおいて，音声の構造的表象を利用する手法を提案した．日本語の大語彙音声認識実験の結果，HMM/GMM ベースのシステムから 6.69% の CER 削減を実現することができた．

今後の課題として，識別的リランキングの feature engineering がある．構造的表象に限らず，例えば単語 n-gram カウントのような言語的特徴量も，識別的リランキングには有効であることが知られている．これらは異なる種類の情報を捉えていると考えられるため，同時に利用すれば，音声認識率の向上に有効ではないかと考えられる．

第6章

まとめ

6.1　まとめ

音声認識は様々な応用に使える技術であり，これまで様々な手法が提案されてきた．具体的には例えば，VAD で音声区間を求め，その区間から求めた MFCC や PLP などの音響特徴量を，VTLN し，それを前後数フレーム連結して LDA し，STC し，fMLLR し，特徴量強調し，特徴量正規化したものを特徴量にして，音響モデルとして HMM/DNN，言語モデルとして modified Kneser-Ney smoothing をかけた N-gram を用いて WFST デコーダで音声認識し，それを様々な特徴量を利用した識別モデルでリランキングし，このようなシステムを複数集めてシステムコンビネーションすることが行われる．

本論文では，上記のような多岐に渡る音声認識の要素技術の中で，今後も利用されていくであろう重要な技術を見つけ，それを改良するための手法を提案した．まず，高速に動作する非定常雑音環境下に頑健な特徴量強調手法として，SPLICE や NMN-SPLICE を越える精度を持つ，DPLT（Discriminative Piecewise Linear Transformation）を提案した．加えて，ミスマッチがない場合に精度を低下を引き起こさないために，クリーン音声とノイジー音声を識別する CN 識別の導入も提案した．次に，音声認識結果を識別モデルでリランキングする際に，時間的に離れた部分の音響的な関連性をとらえた特徴量である音声の構造的表象を特徴量として利用する手法を提案し，音声認識システムのさらなる精度向上を実現した．

6.2　今後の展望

第 5 章で扱った識別的リランキングにおける特徴量に関する研究は，まだ研究の余地がある．本論文では特徴量として音声の構造的表象のみを利用したが，これと他の特徴量を組み合わせることが考えられる [70]．今回扱った長時間に渡る音響的な特徴量や，他にも韻律的な特徴量，言語的特徴量など含め，様々な特徴量をすべて同時に利用することにより，どの程度精度が向上するのか，今後見極めていく必要がある．また，本論文では，識別的リランキングのモデルとして，N-best リストを平均化パーセプトロンでリランキングするモデルを利用したが，入力をラティスにしたり，モデルを複雑なものにすることで，さらなる精度向上が見込まれる．

また他に，音響モデルとして HMM/DNN が利用されるようになってから比較的日が浅いため，HMM/DNN に合うような特徴量抽出手法を検討していく必要がある．今回提案した DPLT は，HMM/DNN システムで評価を行なえていないため，今後評価を行う必要がある．例えば，HMM/DNN を用いることにより，LDA や fMPE などの特徴量空間の識別学習を利用することによる精度向上がほぼなくなったと報告されているなど，今後，どのような技術を HMM/DNN と組み合わせていくべきかを考えていく必要がある．また，HMM/DNN を用いる場合には，MFCC を使うより，よりシンプルな FBANK を利用した

方が精度が高いという報告もあり [71] 特徴量としてはよりシンプルなものを用いた方がよいという結論が得られるかもしれない．今後の研究が期待される．

謝辞

本研究を進めるにあたり，常日頃からご指導，ご鞭撻を賜りました指導教員の峯松信明教授に深く感謝致します．峯松先生には，学部四年生から博士課程修了までの六年間の長きに渡り，熱心にご指導頂きました．本当にありがとうございました．広瀬啓吉教授にも，六年間の長きに渡りお世話になり，数々の鋭いご指摘ご助言を頂きました．齋藤大輔助教には，日頃から白熱した議論をさせて頂きました．NTT CS 研の吉岡拓也氏，MERL の渡部晋治氏には，特に第 3 章の研究を進める上でお世話になりました．的確な論理思考で研究を組み立てていく姿には圧倒されるばかりでした．IBM 東京基礎研究所の倉田岳人氏，西村雅史氏には，特に第 5 章の研究を進める上でお世話になりました．お二人に出会って研究をさせていただいたことは，私の今後の人生に大きな影響を与えました．ここに感謝の意を申し上げます．

広瀬・峯松研究室の皆様のおかげで有意義な研究生活を送ることができました．論文の共著者として一緒に研究をしてくださった喬宇さん，平野宏子さん，朝川智さん，黒岩龍さん，羅徳安さん，馬学彬さん，印南圭祐さん，國越晶さん，高澤真章君，清水信哉君，千々岩圭吾君，中村綾乃さん，甲斐常伸君，柏木陽佑君，加藤集平君，橋本浩弥君，Teiseki Ou さん，Tongmu Zhao 君，Luan Yi さん，池島純君，黒滝夏子さん，正木大介君，Nguyen Duc Duy 君，岡安貴大君，中村新芽君，槇佑馬君（学年・五十音順）に感謝します．おかげ様で，音声に関する非常に沢山の分野の研究を行うことができました．また研究室活動を生活面から支えてくださった，高橋登技官，秘書の池上恵さんに感謝します．その他の研究室の皆様にも，いろんな場面でお世話になり，感謝しています．

最後に，生まれてから今日まで私を支えてくれた家族に感謝します．また，日頃の生活を支えてくれた，妻の智子に感謝します．

2013 年 1 月 28 日

鈴木 雅之

付録 A

実験結果の詳細

A.1　実験結果

第 3 章 の 表 3.1，第 4 章 の 表 4.1 で結果を示した実験に関して，それぞれの雑音環境，SN での結果を示す．雑音環境は，ノイズクローズドな A セットでは Subway，Babble，Car，Exhibition の 4 種類，ノイズオープンな B セットでは Restaurant，Street，Airport，Station の 4 種類，チャネルノイズを含む C セットでは，Subway，Street の 2 種類が使われている．それぞれの表で，雑音環境は，それぞれの環境の頭 3 文字で示されている．また，全 SN での平均は，SN が 20,15,10,5,0 のもののみの平均であり，∞ と -5 のものは含まれていない．

表 A.1: 特徴量強調なしの結果

clean 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.46	0.39	0.60	0.31	0.44	0.46	0.39	0.60	0.31	0.44	0.46	0.36	0.41
20	2.39	1.00	1.76	2.59	1.94	1.11	1.60	1.25	1.33	1.32	1.87	1.87	1.87
15	5.89	4.23	4.86	7.13	5.53	3.35	4.53	3.31	3.76	3.74	5.99	5.38	5.68
10	19.44	13.72	19.06	22.06	18.57	11.51	16.54	10.17	13.61	12.96	19.25	18.17	18.71
5	49.77	40.99	55.95	54.64	50.34	34.54	45.07	34.36	43.51	39.37	51.09	45.62	48.36
0	75.65	74.43	78.35	78.56	76.75	67.64	74.85	67.40	74.05	70.98	76.30	74.91	75.60
-5	85.02	86.61	85.51	88.52	86.41	86.12	86.19	82.91	85.16	85.09	84.37	86.09	85.23
平均	30.63	26.87	32.00	33.00	30.62	23.63	28.52	23.30	27.25	25.67	30.90	29.19	30.04

multi 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.58	0.48	0.69	0.46	0.55	0.58	0.48	0.69	0.46	0.55	0.64	0.48	0.56
20	1.11	0.94	0.81	1.27	1.03	1.17	1.48	1.22	0.96	1.21	0.89	1.63	1.26
15	1.60	1.57	1.46	1.48	1.53	1.47	2.09	1.37	1.73	1.67	1.84	1.84	1.84
10	2.64	2.51	2.68	3.18	2.75	3.10	3.66	2.65	2.93	3.09	3.04	3.57	3.30
5	6.23	6.74	7.49	8.21	7.17	7.77	7.95	6.62	9.01	7.84	6.54	8.80	7.67
0	20.82	28.20	34.09	23.05	26.54	25.70	26.81	22.07	29.34	25.98	20.69	26.42	23.55
-5	63.95	69.14	77.39	60.17	67.66	61.81	68.14	61.86	72.97	66.19	62.70	67.14	64.92
平均	6.48	7.99	9.31	7.44	7.80	7.84	8.40	6.79	8.79	7.96	6.60	8.45	7.53

表 A.2: SPLICE の結果

clean 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.68	0.60	0.72	0.49	0.62	0.68	0.60	0.72	0.49	0.62	0.68	0.63	0.66
20	0.92	0.70	0.69	0.93	0.81	0.71	1.12	0.86	0.77	0.86	1.04	1.03	1.04
15	1.75	1.24	1.13	1.76	1.47	1.35	1.69	0.95	1.57	1.39	2.30	2.48	2.39
10	3.44	2.84	3.28	4.38	3.48	3.47	5.47	2.59	4.41	3.98	5.22	7.38	6.30
5	9.03	12.30	12.56	11.88	11.44	11.48	16.60	10.77	16.41	13.81	12.07	20.98	16.52
0	28.34	42.74	40.83	33.88	36.45	35.74	47.25	36.06	48.29	41.84	35.71	52.39	44.05
-5	63.80	77.03	78.08	68.71	71.91	70.92	78.60	72.92	81.43	75.97	70.03	79.72	74.88
平均	8.70	11.96	11.70	10.57	10.73	10.55	14.43	10.25	14.29	12.38	11.27	16.85	14.06

multi 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.83	0.54	0.57	0.46	0.60	0.83	0.54	0.57	0.46	0.60	0.77	0.54	0.66
20	0.77	0.76	0.78	0.93	0.81	1.38	1.24	1.64	1.39	1.41	1.47	1.12	1.30
15	1.23	1.12	1.13	1.36	1.21	2.98	1.87	2.74	2.56	2.54	2.06	1.84	1.95
10	2.61	2.48	2.30	2.59	2.50	5.65	4.35	5.61	5.62	5.31	2.98	3.93	3.46
5	5.00	7.98	6.08	6.66	6.43	14.25	11.19	11.45	13.85	12.68	7.92	11.61	9.77
0	16.73	31.05	24.37	19.38	22.88	35.62	32.16	29.97	37.43	33.80	23.76	34.28	29.02
-5	50.45	68.35	66.21	53.22	59.56	70.77	66.57	65.08	73.99	69.10	60.12	71.10	65.61
平均	5.27	8.68	6.93	6.18	6.77	11.98	10.16	10.28	12.17	11.15	7.64	10.56	9.10

表 A.3: 区分 $\boldsymbol{y}_t$，線形変換 $\boldsymbol{e}_t(1)$，$\lambda = 0$ の結果

clean 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	9.09	8.31	8.68	9.41	8.87	9.09	8.31	8.68	9.41	8.87	9.86	8.59	9.23
20	0.71	0.60	0.81	0.96	0.77	0.71	0.82	0.95	0.62	0.77	1.38	0.94	1.16
15	1.44	1.03	1.01	1.60	1.27	1.14	1.57	1.07	1.51	1.32	2.39	2.15	2.27
10	2.92	2.51	2.89	3.58	2.98	2.89	4.47	2.45	3.98	3.45	4.33	5.68	5.01
5	7.28	10.04	9.72	10.09	9.28	9.95	13.15	9.31	13.67	11.52	10.90	16.78	13.84
0	25.12	37.42	36.59	29.96	32.27	31.53	41.32	31.61	42.86	36.83	32.51	45.83	39.17
-5	60.76	74.91	75.28	65.54	69.12	69.70	75.91	69.49	78.80	73.48	67.98	77.15	72.56
平均	7.49	10.32	10.20	9.24	9.31	9.24	12.27	9.08	12.53	10.78	10.30	14.28	12.29

multi 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	1.87	1.45	1.49	1.48	1.57	1.87	1.45	1.49	1.48	1.57	2.00	1.66	1.83
20	0.80	0.70	1.01	1.39	0.98	1.20	1.18	1.82	1.51	1.43	3.81	1.60	2.70
15	1.01	0.94	0.84	1.42	1.05	3.10	1.78	2.86	3.27	2.75	3.87	1.93	2.90
10	1.87	2.09	2.12	2.34	2.11	5.37	3.60	5.22	6.05	5.06	3.78	3.51	3.65
5	3.87	6.65	5.43	5.25	5.30	12.93	9.16	11.33	12.74	11.54	8.93	9.73	9.33
0	14.43	27.15	21.15	16.63	19.84	31.84	27.78	27.65	34.50	30.44	23.30	28.69	25.99
-5	46.98	67.41	63.50	51.28	57.29	71.42	64.27	63.94	75.35	68.75	57.57	67.02	62.30
平均	4.40	7.51	6.11	5.41	5.85	10.89	8.70	9.78	11.61	10.24	8.74	9.09	8.92

表 A.4: 区分 $\boldsymbol{y}_t$，線形変換 $\boldsymbol{e}_t(9)$，$\lambda = 10^{-3}$ の結果

clean 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	4.76	4.59	4.74	5.52	4.90	4.76	4.59	4.74	5.52	4.90	6.08	5.71	5.89
20	0.64	0.67	1.43	1.76	1.13	0.61	0.91	0.95	0.68	0.79	1.38	0.91	1.14
15	1.26	0.91	1.64	1.73	1.39	1.11	1.42	0.98	1.48	1.25	2.27	1.51	1.89
10	2.86	2.48	3.01	3.73	3.02	2.64	3.72	2.56	3.18	3.03	4.24	4.26	4.25
5	6.79	8.52	9.90	8.98	8.55	9.06	11.67	8.26	11.23	10.06	10.44	15.15	12.80
0	23.86	34.13	32.96	28.42	29.84	29.69	37.70	29.05	39.28	33.93	32.30	42.08	37.19
-5	59.35	72.64	73.19	63.75	67.23	68.53	74.52	68.09	76.86	72.00	67.30	75.18	71.24
平均	7.08	9.34	9.79	8.92	8.78	8.62	11.08	8.36	11.17	9.81	10.13	12.78	11.45

multi 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	1.32	1.15	1.28	1.08	1.21	1.32	1.15	1.28	1.08	1.21	1.44	1.39	1.41
20	0.68	0.97	1.70	1.51	1.22	0.89	0.97	1.46	1.17	1.12	4.70	1.06	2.88
15	0.89	0.85	1.01	1.36	1.03	2.52	1.42	2.03	2.28	2.06	4.05	1.60	2.82
10	1.81	1.90	1.85	2.44	2.00	4.64	3.02	3.76	4.72	4.04	4.79	3.33	4.06
5	3.32	5.56	5.52	4.94	4.83	10.50	7.22	9.36	10.89	9.49	9.24	8.31	8.77
0	13.72	24.24	19.80	17.31	18.77	29.66	25.27	24.87	31.63	27.86	23.27	27.03	25.15
-5	45.47	66.05	60.90	49.80	55.55	69.14	62.27	63.91	73.96	67.32	56.16	66.78	61.47
平均	4.08	6.70	5.98	5.51	5.57	9.64	7.58	8.30	10.14	8.91	9.21	8.27	8.74

表 A.5: NMN-SPLICE の結果

clean 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	1.29	1.15	1.43	1.05	1.23	1.29	1.15	1.43	1.05	1.23	1.32	1.03	1.17
20	0.92	0.70	0.60	0.93	0.79	0.80	1.03	0.78	0.56	0.79	1.01	1.24	1.12
15	1.66	1.42	1.10	1.82	1.50	1.29	1.63	1.16	1.54	1.41	1.87	2.42	2.15
10	4.14	2.81	3.13	3.70	3.45	3.10	4.69	2.68	3.21	3.42	4.42	5.02	4.72
5	9.86	9.85	9.75	10.95	10.10	9.36	10.58	9.25	10.89	10.02	10.10	12.15	11.12
0	29.60	36.31	41.34	30.89	34.53	30.98	37.73	32.51	39.86	35.27	31.07	35.25	33.16
-5	67.42	76.57	81.99	65.44	72.86	71.11	75.42	72.71	79.98	74.81	68.10	75.85	71.97
平均	9.24	10.22	11.18	9.66	10.07	9.11	11.13	9.28	11.21	10.18	9.69	11.22	10.46

multi 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.71	0.79	0.92	0.68	0.78	0.71	0.79	0.92	0.68	0.78	0.64	0.73	0.69
20	0.83	0.76	0.81	0.86	0.82	1.04	1.45	1.34	0.93	1.19	2.18	2.09	2.14
15	1.14	1.27	1.13	1.30	1.21	2.33	1.72	1.67	1.64	1.84	2.27	2.21	2.24
10	2.30	2.42	2.30	2.59	2.40	4.30	3.51	3.01	3.27	3.52	2.67	3.23	2.95
5	5.25	6.02	5.61	6.51	5.85	9.33	6.86	8.05	8.98	8.30	5.34	7.35	6.34
0	17.41	26.54	27.86	19.38	22.80	25.85	25.03	24.93	30.61	26.61	15.84	22.40	19.12
-5	49.40	65.66	72.59	46.96	58.65	65.83	60.46	63.47	71.95	65.43	47.47	59.58	53.52
平均	5.39	7.40	7.54	6.13	6.61	8.57	7.71	7.80	9.09	8.29	5.66	7.46	6.56

表 A.6: 区分 $\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$，線形変換 $\boldsymbol{y}_t$，$\lambda = 0$ の結果

clean 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.49	0.45	0.69	0.43	0.52	0.49	0.45	0.69	0.43	0.52	0.55	0.48	0.52
20	0.71	0.73	0.69	1.14	0.82	0.80	0.97	0.72	0.68	0.79	0.98	1.06	1.02
15	1.57	1.39	1.13	1.67	1.44	1.23	1.81	1.10	1.60	1.43	1.81	1.90	1.86
10	3.75	2.87	3.37	3.70	3.42	3.04	4.38	2.36	3.58	3.34	4.11	4.99	4.55
5	10.04	9.89	11.18	10.80	10.48	9.55	12.39	9.60	11.17	10.68	10.72	14.84	12.78
0	30.40	37.55	40.59	32.74	35.32	31.41	39.06	31.73	39.25	35.36	33.74	42.78	38.26
-5	63.74	76.06	79.78	65.47	71.26	71.38	75.57	70.98	77.60	73.88	67.67	77.57	72.62
平均	9.29	10.49	11.39	10.01	10.30	9.21	11.72	9.10	11.26	10.32	10.27	13.11	11.69

multi 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.74	0.70	0.78	0.80	0.75	0.74	0.70	0.78	0.80	0.75	0.64	0.63	0.63
20	0.64	0.70	0.86	1.02	0.80	1.04	1.09	1.19	0.83	1.04	1.23	1.27	1.25
15	1.23	1.18	1.04	1.51	1.24	1.84	1.57	1.64	1.45	1.62	1.54	1.90	1.72
10	2.15	2.36	2.30	2.96	2.44	3.47	3.23	3.37	3.27	3.34	2.67	3.36	3.02
5	5.65	7.16	6.02	6.66	6.37	9.58	8.22	8.71	9.16	8.92	6.39	8.65	7.52
0	18.58	29.14	29.05	19.69	24.11	27.48	27.00	26.42	32.58	28.37	19.71	28.69	24.20
-5	51.24	69.07	75.66	49.21	61.30	68.71	64.18	64.93	74.70	68.13	55.08	68.56	61.82
平均	5.65	8.11	7.85	6.37	7.00	8.68	8.22	8.27	9.46	8.66	6.31	8.77	7.54

表 A.7: 区分 $\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$，線形変換 $\boldsymbol{e}_t(1)$，$\lambda = 0$ の結果

clean 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.58	0.88	0.89	0.77	0.78	0.58	0.88	0.89	0.77	0.78	0.74	0.85	0.79
20	0.74	0.70	0.69	1.42	0.89	0.83	0.85	0.78	0.56	0.75	0.98	1.03	1.01
15	1.32	1.24	1.19	1.70	1.36	1.01	1.57	1.13	1.36	1.27	1.69	2.09	1.89
10	3.35	2.60	3.31	3.42	3.17	2.58	3.78	2.48	3.24	3.02	4.05	4.75	4.40
5	8.78	8.68	9.48	10.00	9.24	8.69	10.37	8.23	10.09	9.35	9.76	12.79	11.28
0	27.76	34.40	37.55	28.82	32.13	29.44	35.49	29.08	37.58	32.90	29.17	37.73	33.45
-5	61.71	73.70	77.78	61.43	68.66	68.96	73.00	68.57	77.20	71.93	65.61	75.97	70.79
平均	8.39	9.52	10.44	9.07	9.36	8.51	10.41	8.34	10.57	9.46	9.13	11.68	10.40

multi 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.64	0.79	0.84	0.89	0.79	0.64	0.79	0.84	0.89	0.79	0.80	0.88	0.84
20	0.83	0.82	1.16	1.42	1.06	1.29	1.27	1.28	0.89	1.18	1.20	1.54	1.37
15	1.01	1.21	1.01	1.94	1.29	1.78	1.72	1.58	1.64	1.68	1.44	1.84	1.64
10	1.90	2.09	2.21	2.93	2.28	3.56	2.99	3.04	3.15	3.19	2.27	3.30	2.78
5	4.61	5.83	5.76	6.05	5.56	9.46	7.04	7.72	8.73	8.24	5.04	7.86	6.45
0	17.22	26.72	26.78	17.71	22.11	25.97	24.30	24.22	31.01	26.38	17.29	24.79	21.04
-5	48.51	68.20	72.83	47.33	59.22	66.93	62.42	63.85	74.64	66.96	52.10	66.20	59.15
平均	5.11	7.33	7.38	6.01	6.46	8.41	7.46	7.57	9.08	8.13	5.45	7.87	6.66

表 A.8: 区分 $\boldsymbol{y}_t - \hat{\boldsymbol{n}}_t$，線形変換 $\boldsymbol{e}_t(9)$，$\lambda = 10^{-3}$ の結果

clean 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	3.56	3.17	3.37	3.24	3.34	3.56	3.17	3.37	3.24	3.34	4.97	4.59	4.78
20	0.64	0.70	0.78	0.93	0.76	0.71	0.82	0.66	0.62	0.70	0.61	0.97	0.79
15	1.20	1.24	0.78	1.33	1.14	1.20	1.33	1.04	1.17	1.19	1.54	2.06	1.80
10	2.30	2.39	2.65	2.81	2.54	2.61	3.51	2.09	3.09	2.83	3.01	4.11	3.56
5	6.54	7.62	7.22	8.30	7.42	7.74	8.74	7.07	8.76	8.08	8.01	11.37	9.69
0	22.81	31.95	32.39	25.08	28.06	27.05	30.99	26.57	33.69	29.57	25.64	33.74	29.69
-5	56.59	72.79	75.10	56.59	65.27	67.82	70.04	67.34	74.67	69.97	60.42	73.67	67.05
平均	6.70	8.78	8.76	7.69	7.98	7.86	9.08	7.49	9.47	8.47	7.76	10.45	9.11

multi 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	1.04	0.94	0.89	0.77	0.91	1.04	0.94	0.89	0.77	0.91	1.20	1.06	1.13
20	0.71	0.79	0.72	0.62	0.71	1.17	1.09	1.04	0.77	1.02	0.98	1.30	1.14
15	1.01	0.97	0.75	1.33	1.02	1.93	1.21	1.37	1.33	1.46	1.38	1.93	1.66
10	1.78	2.18	1.88	2.50	2.09	3.50	2.63	2.39	3.12	2.91	1.93	3.14	2.54
5	4.24	6.38	5.46	5.55	5.41	9.12	6.38	7.25	7.96	7.68	5.68	7.92	6.80
0	16.58	26.42	24.34	18.11	21.36	25.67	23.94	23.62	30.05	25.82	16.58	24.33	20.45
-5	46.91	67.50	71.40	46.96	58.19	67.12	62.48	63.38	73.06	66.51	50.11	65.45	57.78
平均	4.86	7.35	6.63	5.62	6.12	8.28	7.05	7.13	8.65	7.78	5.31	7.72	6.52

表 A.9: 区分 $\boldsymbol{v}_t(1)$，線形変換 $\boldsymbol{y}_t$，$\lambda = 0$ の結果

clean 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.52	0.45	0.57	0.37	0.48	0.52	0.45	0.57	0.37	0.48	0.52	0.48	0.50
20	0.68	0.57	0.72	0.89	0.71	0.64	1.00	0.72	0.65	0.75	0.77	0.97	0.87
15	1.54	1.12	0.98	1.45	1.27	0.89	1.42	0.86	1.23	1.10	1.50	1.48	1.49
10	2.89	2.42	2.68	3.33	2.83	2.89	3.81	2.12	3.39	3.05	3.62	5.86	4.74
5	8.66	8.65	9.90	9.75	9.24	8.04	11.64	7.93	11.14	9.69	11.70	17.53	14.62
0	26.04	32.92	36.03	28.57	30.89	27.11	38.48	27.14	37.89	32.66	36.63	49.09	42.86
-5	61.41	71.16	74.62	62.05	67.31	66.47	74.52	66.33	75.50	70.70	71.02	78.72	74.87
平均	7.96	9.14	10.06	8.80	8.99	7.91	11.27	7.75	10.86	9.45	10.84	14.99	12.92

multi 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.71	0.60	0.92	0.59	0.70	0.71	0.60	0.92	0.59	0.70	0.68	0.54	0.61
20	0.68	0.76	0.84	1.17	0.86	0.71	1.30	1.43	0.77	1.05	0.89	1.27	1.08
15	1.01	1.06	1.07	1.64	1.20	1.60	1.72	1.79	1.73	1.71	1.23	1.51	1.37
10	2.33	1.78	2.39	3.09	2.40	3.19	3.57	3.91	3.98	3.66	2.76	4.08	3.42
5	5.89	6.92	7.07	7.00	6.72	8.66	9.01	9.63	10.46	9.44	7.89	12.24	10.06
0	18.45	27.33	26.93	20.21	23.23	26.87	29.56	24.81	32.18	28.36	26.25	38.54	32.39
-5	49.00	65.57	68.80	50.94	58.58	68.07	65.27	62.33	70.69	66.59	64.75	73.19	68.97
平均	5.67	7.57	7.66	6.62	6.88	8.21	9.03	8.31	9.82	8.84	7.80	11.53	9.67

表 A.10: 区分 $\boldsymbol{v}_t(1)$，線形変換 $\boldsymbol{e}_t(1)$，$\lambda = 0$ の結果

clean 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.77	0.85	0.95	0.56	0.78	0.77	0.85	0.95	0.56	0.78	0.86	0.91	0.88
20	0.61	0.63	0.86	1.73	0.96	0.74	0.88	0.75	0.77	0.78	0.89	1.09	0.99
15	1.26	1.06	1.16	1.91	1.35	0.89	1.33	0.89	1.08	1.05	1.35	2.03	1.69
10	2.36	2.33	2.51	3.67	2.72	2.55	3.30	1.82	3.12	2.70	3.53	5.05	4.29
5	7.03	7.71	8.59	8.98	8.08	7.34	9.82	7.07	9.87	8.52	10.35	14.42	12.38
0	23.12	29.93	32.57	26.13	27.94	25.36	33.65	24.90	34.00	29.48	32.73	43.56	38.15
-5	57.72	68.98	72.80	59.02	64.63	64.60	71.86	63.47	73.50	68.36	68.41	76.78	72.59
平均	6.88	8.33	9.14	8.48	8.21	7.38	9.80	7.09	9.77	8.51	9.77	13.23	11.50

multi 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.71	0.67	0.81	0.43	0.66	0.71	0.67	0.81	0.43	0.66	0.74	0.57	0.65
20	0.83	0.63	0.92	2.10	1.12	0.86	1.18	1.16	1.02	1.05	0.86	1.72	1.29
15	1.14	0.88	1.28	2.07	1.34	1.66	1.45	1.49	1.76	1.59	1.35	2.12	1.73
10	1.84	1.78	2.45	3.70	2.44	2.98	2.84	3.67	3.64	3.28	2.43	3.81	3.12
5	5.25	6.32	6.77	7.31	6.41	8.29	7.47	8.89	10.43	8.77	6.32	10.31	8.31
0	15.93	25.39	25.35	18.94	21.40	25.48	26.66	22.96	30.18	26.32	23.46	33.80	28.63
-5	47.28	65.48	69.43	49.49	57.92	68.07	63.66	60.75	70.93	65.85	62.33	70.50	66.41
平均	5.00	7.00	7.35	6.82	6.54	7.85	7.92	7.63	9.41	8.20	6.88	10.35	8.62

表 A.11: 区分 $\boldsymbol{v}_t(1)$，線形変換 $\boldsymbol{e}_t(9)$，$\lambda = 10^{-3}$ の結果

clean 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.77	1.09	1.25	0.77	0.97	0.77	1.09	1.25	0.77	0.97	1.04	1.33	1.19
20	0.61	0.60	0.78	1.45	0.86	0.71	0.88	0.75	0.43	0.69	0.61	0.97	0.79
15	1.11	0.91	1.22	1.67	1.23	0.77	1.15	0.69	0.89	0.88	0.98	1.45	1.22
10	1.81	2.06	2.48	2.99	2.34	2.15	2.81	1.91	2.41	2.32	2.64	3.99	3.31
5	5.74	7.13	7.75	7.16	6.94	6.60	7.80	6.41	8.24	7.26	8.78	12.52	10.65
0	20.33	27.63	28.42	23.23	24.90	23.18	30.08	22.19	30.33	26.45	29.08	40.11	34.59
-5	53.48	66.54	69.79	56.09	61.47	62.63	68.86	61.23	70.75	65.87	65.67	74.85	70.26
平均	5.92	7.67	8.13	7.30	7.25	6.68	8.54	6.39	8.46	7.52	8.42	11.81	10.11

multi 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.64	0.57	0.78	0.62	0.65	0.64	0.57	0.78	0.62	0.65	0.68	0.60	0.64
20	0.71	0.60	0.89	1.36	0.89	0.71	0.94	0.95	0.77	0.84	0.74	0.94	0.84
15	0.89	0.88	0.78	1.45	1.00	1.35	1.18	1.28	1.42	1.31	1.11	1.63	1.37
10	1.50	1.81	1.94	2.78	2.01	2.67	2.21	2.65	3.39	2.73	2.09	3.26	2.68
5	4.24	5.99	5.79	5.43	5.36	7.06	6.23	7.61	9.04	7.49	5.80	9.28	7.54
0	13.79	23.00	21.44	16.91	18.79	23.92	24.00	20.61	28.17	24.18	20.02	30.93	25.48
-5	42.92	63.75	65.20	47.82	54.92	64.51	60.10	59.44	68.84	63.22	57.23	69.01	63.12
平均	4.23	6.46	6.17	5.59	5.61	7.14	6.91	6.62	8.56	7.31	5.95	9.21	7.58

表 A.12: 区分 $\boldsymbol{v}_t(9)$，線形変換 $\boldsymbol{y}_t$，$\lambda = 0$ の結果

clean 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.52	0.42	0.66	0.31	0.48	0.52	0.42	0.66	0.31	0.48	0.55	0.42	0.48
20	0.80	0.51	0.75	0.77	0.71	0.64	0.97	0.84	0.62	0.77	0.55	0.88	0.71
15	1.57	0.91	1.01	1.05	1.13	1.04	1.48	0.89	1.20	1.15	1.14	1.39	1.27
10	2.61	2.27	2.18	3.05	2.53	2.52	2.78	2.00	2.75	2.51	3.04	4.11	3.58
5	6.23	7.83	7.40	8.27	7.43	7.37	9.04	6.80	9.29	8.13	9.36	13.60	11.48
0	22.66	30.23	30.96	25.18	27.26	24.16	32.32	23.53	33.85	28.46	31.10	42.90	37.00
-5	57.78	68.98	74.23	60.17	65.29	63.19	72.01	64.72	74.64	68.64	68.53	78.23	73.38
平均	6.77	8.35	8.46	7.66	7.81	7.15	9.32	6.81	9.54	8.20	9.04	12.58	10.81

multi 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.68	0.42	0.72	0.43	0.56	0.68	0.42	0.72	0.43	0.56	0.71	0.42	0.56
20	0.83	0.82	0.86	0.86	0.84	0.86	1.06	1.34	0.62	0.97	0.77	1.45	1.11
15	1.47	0.97	1.16	1.20	1.20	1.84	1.57	1.67	1.91	1.75	1.23	1.84	1.54
10	2.46	2.06	2.33	3.15	2.50	3.32	2.84	3.43	3.46	3.26	2.67	3.81	3.24
5	5.65	7.16	7.40	7.00	6.80	9.03	8.59	8.80	9.97	9.10	7.43	11.94	9.69
0	19.56	28.87	28.51	21.97	24.73	26.04	30.35	24.87	32.55	28.45	26.90	39.45	33.18
-5	53.45	67.65	75.25	58.22	63.64	66.53	69.68	65.11	75.04	69.09	67.21	76.09	71.65
平均	5.99	7.98	8.05	6.84	7.21	8.22	8.88	8.02	9.70	8.71	7.80	11.70	9.75

表 A.13: 区分 $\boldsymbol{v}_t(9)$，線形変換 $\boldsymbol{e}_t(1)$，$\lambda = 0$ の結果

clean 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.80	0.79	0.89	0.59	0.77	0.80	0.79	0.89	0.59	0.77	0.86	0.85	0.85
20	0.68	0.57	0.81	0.68	0.69	0.77	0.97	0.89	0.59	0.80	0.68	1.42	1.05
15	1.35	0.94	0.95	1.17	1.10	0.89	1.33	0.84	1.08	1.04	1.23	2.21	1.72
10	2.09	2.21	1.94	2.78	2.25	2.39	2.39	1.76	2.72	2.32	2.76	4.11	3.44
5	5.40	7.32	6.38	7.31	6.60	6.94	7.92	6.86	8.73	7.61	8.29	12.18	10.24
0	21.19	28.05	28.54	23.54	25.33	22.41	29.35	21.68	30.73	26.04	28.55	39.42	33.98
-5	55.33	67.62	72.35	58.16	63.37	61.96	69.77	62.00	73.40	66.78	66.23	76.00	71.11
平均	6.14	7.82	7.72	7.10	7.20	6.68	8.39	6.41	8.77	7.56	8.30	11.87	10.09

multi 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.80	0.57	0.84	0.62	0.71	0.80	0.57	0.84	0.62	0.71	0.77	0.60	0.68
20	0.86	0.73	0.78	0.77	0.79	0.98	1.24	0.98	0.74	0.98	0.92	1.63	1.27
15	1.26	0.91	0.86	1.20	1.06	1.66	1.48	1.40	1.51	1.51	1.35	2.36	1.85
10	1.90	1.93	1.94	2.87	2.16	2.86	2.57	2.71	3.21	2.84	2.30	3.93	3.12
5	4.67	6.41	5.79	6.11	5.75	8.17	7.35	8.14	9.38	8.26	6.26	10.88	8.57
0	17.72	26.60	26.04	20.15	22.63	24.38	27.39	22.96	30.42	26.29	23.67	35.49	29.58
-5	51.03	66.17	71.88	55.14	61.05	64.14	66.35	62.57	72.85	66.48	62.88	73.40	68.14
平均	5.28	7.32	7.08	6.22	6.48	7.61	8.01	7.24	9.05	7.98	6.90	10.86	8.88

表 A.14: 区分 $\boldsymbol{v}_t(9)$，線形変換 $\boldsymbol{e}_t(9)$，$\lambda = 10^{-3}$ の結果

clean 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.92	0.79	0.84	0.59	0.79	0.92	0.79	0.84	0.59	0.79	0.89	0.97	0.93
20	0.52	0.60	0.84	0.74	0.67	0.68	1.00	0.89	0.46	0.76	0.58	1.27	0.92
15	0.89	0.76	0.75	1.27	0.92	0.89	1.39	1.07	0.99	1.08	1.04	1.90	1.47
10	1.87	1.90	2.03	2.78	2.15	2.43	2.51	1.82	2.44	2.30	2.55	4.20	3.38
5	4.94	6.38	6.20	6.42	5.98	6.66	7.10	6.23	7.53	6.88	6.82	12.52	9.67
0	18.08	26.51	26.07	20.43	22.77	22.04	27.96	20.82	28.97	24.95	25.12	38.24	31.68
-5	51.27	65.84	70.09	56.16	60.84	60.55	67.71	60.42	71.52	65.05	62.45	74.67	68.56
平均	5.26	7.23	7.18	6.33	6.50	6.54	7.99	6.17	8.08	7.19	7.22	11.63	9.42

multi 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.74	0.76	0.75	0.43	0.67	0.74	0.76	0.75	0.43	0.67	0.71	0.79	0.75
20	0.68	0.73	0.72	0.56	0.67	0.64	0.94	0.98	0.56	0.78	0.86	1.15	1.01
15	0.95	0.79	0.66	1.08	0.87	1.35	1.60	1.19	1.27	1.35	1.26	1.75	1.51
10	1.90	1.57	1.58	2.44	1.87	2.70	2.21	2.24	2.75	2.47	1.84	3.42	2.63
5	4.24	5.77	5.67	5.89	5.39	7.15	6.71	7.43	7.84	7.28	5.62	10.31	7.97
0	16.00	25.42	25.11	19.41	21.48	22.84	25.73	21.68	29.53	24.95	22.20	33.28	27.74
-5	48.23	65.33	70.59	53.26	59.35	62.94	63.94	61.11	71.27	64.81	60.33	72.01	66.17
平均	4.75	6.86	6.75	5.88	6.06	6.94	7.44	6.70	8.39	7.37	6.36	9.98	8.17

表 A.15: 表 4.1 の結果の詳細

clean 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.31	0.36	0.48	0.15	0.33	0.31	0.36	0.48	0.15	0.33	0.31	0.36	0.34
20	1.04	1.03	0.86	1.20	1.03	0.77	1.36	0.84	0.99	0.99	1.29	1.60	1.45
15	3.35	3.30	3.19	4.57	3.60	2.15	3.05	2.03	2.90	2.53	3.44	3.23	3.34
10	10.13	11.64	13.99	13.39	12.29	8.66	11.70	7.93	10.46	9.69	9.64	10.55	10.09
5	29.72	33.80	38.98	34.96	34.36	29.01	31.74	27.86	33.42	30.51	26.28	27.09	26.69
0	58.98	63.91	69.88	67.36	65.03	58.98	62.24	57.05	66.24	61.13	53.27	59.10	56.19
-5	80.81	85.04	86.22	85.50	84.39	82.25	83.52	80.14	84.29	82.55	78.60	81.95	80.28
平均	20.64	22.74	25.38	24.30	23.26	19.91	22.02	19.14	22.80	20.97	18.78	20.31	19.55

multi 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.46	0.67	0.63	0.52	0.57	0.46	0.67	0.63	0.52	0.57	0.37	0.57	0.47
20	0.58	0.97	0.81	0.65	0.75	0.64	1.15	0.78	0.56	0.78	0.49	1.00	0.74
15	0.95	1.15	0.89	1.11	1.03	0.86	1.15	1.25	0.96	1.05	0.92	1.18	1.05
10	1.41	2.36	1.55	2.34	1.91	2.43	2.42	1.73	1.94	2.13	1.72	2.75	2.23
5	3.84	7.35	4.53	6.08	5.45	6.72	6.56	5.55	6.20	6.26	4.76	7.44	6.10
0	13.48	25.51	18.52	19.16	19.17	23.24	22.70	18.61	23.48	22.01	17.35	24.94	21.14
-5	46.95	70.22	60.87	51.84	57.47	63.92	62.85	58.25	64.61	62.41	55.02	66.60	60.81
平均	4.05	7.47	5.26	5.87	5.66	6.78	6.80	5.58	6.63	6.45	5.05	7.46	6.25

select 条件

	A セット					B セット					C セット		
SN	Sub.	Bab.	Car	Exh.	平均	Res.	Str.	Air.	Sta.	平均	Sub.	Str.	平均
∞	0.31	0.36	0.51	0.15	0.33	0.31	0.36	0.51	0.15	0.33	0.31	0.36	0.34
20	0.58	0.97	0.81	0.65	0.75	0.61	1.21	0.92	0.59	0.83	0.49	1.00	0.74
15	0.95	1.15	0.89	1.11	1.03	0.86	1.15	1.25	0.96	1.05	0.92	1.18	1.05
10	1.41	2.36	1.55	2.34	1.91	2.43	2.42	1.73	1.94	2.13	1.72	2.75	2.23
5	3.84	7.35	4.53	6.08	5.45	6.72	6.56	5.55	6.20	6.26	4.76	7.44	6.10
0	13.48	25.51	18.52	19.16	19.17	23.24	22.70	18.61	23.48	22.01	17.35	24.94	21.14
-5	46.95	70.22	60.87	51.84	57.47	63.92	62.85	58.25	64.61	62.41	55.02	66.60	60.81
平均	4.05	7.47	5.26	5.87	5.66	6.77	6.81	5.61	6.63	6.46	5.05	7.46	6.25

付録B

区分的線形変換の実装

B.1　区分的線形変換の実装

第 2 章において，ステレオデータを用いた区分的線形変換に関する技術を提案した．ここでは，これの実装方法について述べる．

入力のノイジーな音声の特徴量を $\boldsymbol{y}_t$，求めるべきクリーンな音声の特徴量を $\boldsymbol{x}_t$ とおく．区分的線形変換による特徴量強調の枠組みでは，まず，部分空間のインデックスとして k を導入し，$p(k|\boldsymbol{y}_t)$ を定義する．$p(k|\boldsymbol{y}_t)$ は，SPLICE であれば $\boldsymbol{y}_t$ の GMM を学習することで計算されるし，本論文での提案手法であれば，クリーン音声状態の識別の結果得られる．ここでは $p(k|\boldsymbol{y}_t)$ を計算するプログラムは既に実装できていると仮定する．

次に，$p(k|\boldsymbol{y}_t)$ を元に，線形変換 $\boldsymbol{A}_k$ を学習する．これは，重み付き最小二乗誤差基準を使う場合，以下の数式で定式化される．

$$\underset{\boldsymbol{A}_k}{\operatorname{argmax}} \sum_t p(k|\boldsymbol{y}_t) \, ||\boldsymbol{A}_k \boldsymbol{e}_t - \boldsymbol{x}_t||^2 \tag{B.1}$$

ただしここで $\boldsymbol{e}_t$ は，SPLICE では $\boldsymbol{e}_t = \begin{bmatrix} 1 \\ \boldsymbol{y}_t \end{bmatrix}$，本論文の提案手法であれば，$\boldsymbol{y}_t$ とそこから得られる雑音特徴量の推定値とそれらの前後数フレームを連結したものである．

(B.1) を解くためのプログラムを，matlab 的な記法で書いたものを以下に示す．

```
for ii = 1:nfile
    % ステレオデータを読む
    y = read_feature_file( noisyfiles{ii} );
    x = read_feature_file( cleanfiles{ii} );

    [ndim nframe] = size(Y); % = size(X)

    for t = 1:nframe
        % p(k|y) を計算
        p_of_k_given_y = calc_posterior( y(:,t) );

        e = [1; y(:,t)];
        for k = 1:nmix
            if p_of_k_given_y(k) > 0.0001 % 高速化
                S_xe(:,:,k) = S_xe(:,:,k) + p_of_k_given_y(k) * x(:,t) * e';
                S_ee (:,:,k) = S_ee(:,:,k) + p_of_k_given_y(k) * e * e';
            end
        end
```

```
    end
end

for k = 1:nmix
    A(:,:,k) = S_xe(:,:,k) / S_ee(:,:,k);
end
```

S_{xe} や S_{ee} は，$p(k|\boldsymbol{y}_t)$ で重みを付けられた．入力特徴量 e_t と出力すべき特徴量 $\boldsymbol{x}_t$ の相互共分散行列，$\boldsymbol{e}_t$ の分散共分散行列に相当するものである．求めるべき $\boldsymbol{A}_k$ は，相互共分散行列を，$\boldsymbol{e}_t$ の分散共分散行列で正規化したような形として得られる．

「高速化」と書いた部分では，$p(k|\boldsymbol{y}_t)$ が十分に小さい場合に処理を省くことで，高速化を実現している．特に GMM を用いて事後確率を計算すると，その事後確率はスパースになりやすい傾向がある．本論文の実験では上記に示した通り，$p(k|\boldsymbol{y}_t)$ が 10^{-4} より小さくなる場合に処理を省いた．これにより，GMM の混合数として 1024 を利用していても，実際にこの部分では，10 回程度の計算になる．この高速化を導入しても，$\boldsymbol{A}_k$ の推定結果と最終的な音声認識の精度には，ほぼ影響はない．

なお上記による SPLICE の実装として，著者が Matlab で実装を行った `https://sites.google.com/site/suzukimasayuki/splice` が利用できる．

付録C

正規分布に関する公式

C.1　条件付き正規分布

二つの変数 $\boldsymbol{x}$ と $\boldsymbol{y}$ の結合ベクトルが正規分布に従う時，$\boldsymbol{y}$ が与えられたときの $\boldsymbol{x}$ の条件付き確率分布も正規分布となる．

まず結合ベクトルが以下のような正規分布に従うとする．

$$p\left(\left[\begin{array}{c}\boldsymbol{x}\\ \boldsymbol{y}\end{array}\right]\right) = \mathcal{N}\left(\left[\begin{array}{c}\boldsymbol{x}\\ \boldsymbol{y}\end{array}\right], \left[\begin{array}{c}\boldsymbol{\mu}^{\boldsymbol{x}}\\ \boldsymbol{\mu}^{\boldsymbol{y}}\end{array}\right], \left[\begin{array}{cc}\boldsymbol{\Sigma}^{\boldsymbol{xx}} & \boldsymbol{\Sigma}^{\boldsymbol{xy}}\\ \boldsymbol{\Sigma}^{\boldsymbol{yx}} & \boldsymbol{\Sigma}^{\boldsymbol{yy}}\end{array}\right]\right) \tag{C.1}$$

このとき，$p(\boldsymbol{x}|\boldsymbol{y})$ は以下のような正規分布になる．

$$p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}^{\boldsymbol{x}} + \boldsymbol{\Sigma}^{\boldsymbol{xy}}(\boldsymbol{\Sigma}^{\boldsymbol{yy}})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}^{\boldsymbol{y}}), \boldsymbol{\Sigma}^{\boldsymbol{xx}} - \boldsymbol{\Sigma}^{\boldsymbol{xy}}(\boldsymbol{\Sigma}^{\boldsymbol{yy}})^{-1}\boldsymbol{\Sigma}^{\boldsymbol{yx}}\right) \tag{C.2}$$

この正規分布の平均部分はさらに以下のように計算できる．

$$\begin{aligned}\boldsymbol{\mu}^{\boldsymbol{x}} + \boldsymbol{\Sigma}^{\boldsymbol{xy}}(\boldsymbol{\Sigma}^{\boldsymbol{yy}})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}^{\boldsymbol{y}}) &= \left[\boldsymbol{\mu}^{\boldsymbol{x}} - \boldsymbol{\Sigma}^{\boldsymbol{xy}}(\boldsymbol{\Sigma}^{\boldsymbol{yy}})^{-1}\boldsymbol{\mu}^{\boldsymbol{y}}, \boldsymbol{\Sigma}^{\boldsymbol{xy}}(\boldsymbol{\Sigma}^{\boldsymbol{yy}})^{-1}\right]\left[\begin{array}{c}1\\ \boldsymbol{y}\end{array}\right] &\text{(C.3)}\\ &= \overline{\boldsymbol{X}\boldsymbol{Y}}^{\top}\left(\overline{\boldsymbol{Y}\boldsymbol{Y}}^{\top}\right)^{-1}\left[\begin{array}{c}1\\ \boldsymbol{y}\end{array}\right] &\text{(C.4)}\end{aligned}$$

ただし，$\overline{\boldsymbol{X}}$，$\overline{\boldsymbol{Y}}$ はそれぞれ $\boldsymbol{x}$，$\left[\begin{array}{c}1\\ \boldsymbol{y}\end{array}\right]$ を学習データのフレーム数分並べた行列である．また結合ベクトルが GMM としてモデル化されていても，以上の式と似たような展開を行うことができて，結論は以下のようになる．

$$\overline{\boldsymbol{X}}\boldsymbol{P}_k\overline{\boldsymbol{Y}}^{\top}\left(\overline{\boldsymbol{Y}}\boldsymbol{P}_k\overline{\boldsymbol{Y}}^{\top}\right)^{-1}\left[\begin{array}{c}1\\ \boldsymbol{y}\end{array}\right] \tag{C.5}$$

ただし，$\boldsymbol{P}$ は $p(k|\boldsymbol{x}, \boldsymbol{y})$ を学習データのフレーム数分並べた物を対角成分にもつ対角行列とする．

これを，SPLICE の重み付き二乗誤差最小基準を用いた線形変換の解析解 (2.29) を見比べると，クリーン音声特徴量 $\boldsymbol{x}$ とノイジー音声特徴量 $\boldsymbol{y}$ の結合ベクトルを GMM としてモデル化し，$p(\boldsymbol{x}|\boldsymbol{y})$ を求める特徴量強調アプローチは SPLICE とほとんど同じ区分的線形変換を用いる手法であることが分かる．両者の違いは $\boldsymbol{P}_k$ の定義のみである．

C.2　正規分布の周辺分布

VTS を用いた音響モデル適応や特徴量強調においては，非線形変換を線形変換で近似していたが，その理由は正規分布の周辺分布の計算が解析的に求めるためである．

まず $\boldsymbol{n}$ を，以下の正規分布に従う変数と定義する．

$$p(\boldsymbol{n}) = \mathcal{N}(\boldsymbol{n}; \boldsymbol{\mu}^{\boldsymbol{n}}, \boldsymbol{\Sigma}^{\boldsymbol{n}}) \tag{C.6}$$

VTS 強調の文脈では，$\boldsymbol{n}$ は雑音特徴量を表している．次に，$\boldsymbol{n}$ が与えられた下で，$\boldsymbol{y}$ を以下の $\boldsymbol{n}$ の線形変換を平均とする正規分布に従う変数と定義する．

$$p(\boldsymbol{y}|\boldsymbol{n}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{n} + \boldsymbol{b}, \boldsymbol{\Sigma}^{\boldsymbol{y}}) \tag{C.7}$$

VTS 強調の文脈では，$\boldsymbol{y}$ は観測されたノイジーな特徴量であり，$\boldsymbol{A}$ 及び $\boldsymbol{b}$ は VTS 近似により求められる．

このとき，$p(\boldsymbol{y})$ は以下のような正規分布になる．

$$\begin{aligned}
p(\boldsymbol{y}) &= \int p(\boldsymbol{y}, \boldsymbol{n}) d\boldsymbol{n} && \text{(C.8)} \\
&= \int p(\boldsymbol{y}|\boldsymbol{n}) p(\boldsymbol{n}) d\boldsymbol{n} && \text{(C.9)} \\
&= \int \mathcal{N}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{n} + \boldsymbol{b}, \boldsymbol{\Sigma}^{\boldsymbol{y}}) \mathcal{N}(\boldsymbol{n}; \boldsymbol{\mu}^{\boldsymbol{n}}, \boldsymbol{\Sigma}^{\boldsymbol{n}}) d\boldsymbol{n} && \text{(C.10)} \\
&= \mathcal{N}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{\mu}^{\boldsymbol{n}} + \boldsymbol{b}, \boldsymbol{\Sigma}^{\boldsymbol{y}} + \boldsymbol{A}\boldsymbol{\Sigma}^{\boldsymbol{n}}\boldsymbol{A}^{\top}) && \text{(C.11)}
\end{aligned}$$

(C.10) から (C.11) の導出には，exp の中身を $\boldsymbol{n}$ に関して平方完成した形であれば積分が解析的に解け定数になることを利用すればよい．

参考文献

[1] 鹿野 清宏, 伊藤 克亘, 河原 達也, 武田 一哉, and 山本 幹雄, editors. **音声認識システム**. オーム社, 2001.

[2] 河原 達也 and 荒木 雅弘, editors. **音声対話システム**. オーム社, 2006.

[3] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4):359 – 393, 1999.

[4] G. Saon and J.-T. Chien. Large-vocabulary continuous speech recognition systems: A look at some recent advances. *Signal Processing Magazine, IEEE*, 29(6):18 –33, nov. 2012.

[5] 藤本 雅清. 音声区間検出の基礎と世界的な研究動向，今後の展開. **電子情報通信学会誌**, 95(8):754 –758, aug. 2012.

[6] T. Shinozaki, Yu. Kubota, and S. Furui. Unsupervisec cross-validation adaptation algorithms for improved adaptation performance. In *ICASSP*, pages 4377 –4380, april 2009.

[7] M.J.F. Gales and P.C. Woodland. Mean and variance adaptation within the mllr framework. *Computer Speech and Language*, 10(4):249 – 264, 1996.

[8] D. Kolossa and R. Haeb-Umbach, editors. *Robust speech recognition of uncertain or missing data*. Springer, 2011.

[9] J. Taghia, J. Taghia, N. Mohammadiha, Jinqiu Sang, V. Bouse, and R. Martin. An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments. In *ICASSP*, pages 4640 –4643, may 2011.

[10] M.J.F. Gales and S.J. Young. Robust continuous speech recognition using parallel model combination. *Speech and Audio Processing, IEEE Transactions on*, 4(5):352 –359, sep 1996.

[11] A. Acero, Li. Deng, T. Kristjansson, and J. Zhang. Hmm adaptation using vector taylor series for noisy speech recognition. In *ICSLP*, pages 869 –872, 2000.

[12] M.J.F. Gales. Model-based techniques for noise robust speech recognition. *PhD thesis*, 1996.

[13] R.C. van Dalen and M.J.F. Gales. Extended vts for noise-robust speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):733 –743, may 2011.

[14] Duncan Macho, Laurent Mauuary, Bernhard No, Yan Ming Cheng, Douglas Ealey, Denis Jouvet, Holly Kelleher, David Pearce, and Fabien Saadoun. Evaluation of a noise-robust dsr front-end on aurora databases. In *INTERSPEECH*, 2002.

[15] J.F. Gemmeke, T. Virtanen, and A. Hurmalainen. Exemplar-based sparse representations for noise robust automatic speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2067 –2080, sept. 2011.

[16] Jort F. Gemmeke and Hugo Van hamme. Advances in noise robust digit recognition using hybrid exemplar-based techniques. In *INTERSPEECH*, 2012.

[17] J.C. Segura, A. de la Torre, M.C. Benitez, and A.M. Peinado. Model-Based Compensation of the Additive Noise for Continuous Speech Recognition. Experiments Using the Aurora II Database and Tasks. *EUROSPEECH*, pages 221–224, 2001.

[18] V. Stouten. Robust Automatic Speech Recognition in Time-varying Environments. *PhD thesis*, 2006.

[19] B. Frey, L. Deng, A. Acero, and T. Kristjansson. Algonquin: Iterating laplace ' s method to remove multiple types of acoustic distortion for robust speech recognition. In *Eurospeech*, volume 2, pages 901–904, 2001.

[20] Steven J. Rennie. GRAPHICAL MODELS FOR ROBUST SPEECH RECOGNITION IN ADVERSE ENVIRONMENTS. *PhD thesis*, 2006.

[21] S.J. Rennie, P.L. Dognin, and P. Fousek. Matched-condition robust dynamic noise adaptation. In *ASRU*, pages 137–140. IEEE, 2011.

[22] J. Droppo, Li Deng., and A. Acero. Evaluation of SPLICE on the Aurora 2 and 3 Tasks. *ICSLP*, pages 29–32, 2002.

[23] M. Afify, X. Cui, and Y. Gao. Stereo-based stochastic mapping for robust speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(7):1325 –1334, sept. 2009.

[24] L. Buera, E. Lleida, A. Miguel, A. Ortega, and O. Saz. Cepstral vector normalization based on stereo data for robust speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(3):1098 –1113, march 2007.

[25] H. Liao and M.J.F. Gales. Issues with uncertainty decoding for noise robust automatic speech recognition. *Speech Communication*, 50(4):265 – 277, 2008.

[26] A. de la Torre, A.M. Peinado, J.C. Segura, J.L. Perez-Cordoba, M.C. Benitez, and A.J. Rubio. Histogram equalization of speech representation for robust speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 13(3):355 – 366, may 2005.

[27] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. In *ICSLP*, volume 2, pages 1137 –1140 vol.2, oct 1996.

[28] O. Kalinli, M.L. Seltzer, J. Droppo, and A. Acero. Noise adaptive training for robust automatic speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(8):1889 –1901, nov. 2010.

[29] G. Heigold, H. Ney, R. Schluter, and S. Wiesler. Discriminative training for automatic speech recognition: Modeling, criteria, optimization, implementation, and performance. *Signal Processing Magazine, IEEE*, 29(6):58 –69, nov. 2012.

[30] D. Povey and P.C. Woodland. Minimum phone error and i-smoothing for improved discriminative training. In *ICASSP*, volume 1, pages I–105 –I–108, may 2002.

[31] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah. Boosted mmi for model and feature-space discriminative training. In *ICASSP*, pages 4057 –4060, 31 2008-april 4 2008.

[32] Bing Zhang, S. Matsoukas, and R. Schwartz. Discriminatively trained region dependent feature transforms for speech recognition. In *ICASSP*, volume 1, page I, may 2006.

[33] Andrew Senior, Youngmin Cho, and Jason Weston. Learning improved linear transforms for speech recognition. In *ICASSP*, 2012.

[34] M.J.F. Gales. Semi-tied covariance matrices for hidden markov models. *IEEE Transactions on Speech and Audio Processing*, 7:272–281, 1999.

[35] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig. fmpe: Discriminatively trained features for speech recognition. In *ICASSP*, volume 1, pages 961 – 964, 18-23, 2005.

[36] J. Droppo, M. Mahajan, A. Gunawardana, and A. Acero. How to train a discriminative front end with stochastic gradient descent and maximum mutual information. In *ASRU*, pages 41 –46, nov. 2005.

[37] G. Zweig, P. Nguyen, D. Van Compernolle, K. Demuynck, L. Atlas, P. Clark, G. Sell, M. Wang, F. Sha, H. Hermansky, D. Karakos, A. Jansen, S. Thomas, G.S.V.S. Sivaram, S. Bowman, and J. Kao. Speech recognitionwith segmental conditional random fields: A summary of the jhu clsp 2010 summer workshop. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5044 –5047, may 2011.

[38] M.J.F. Gales, S. Watanabe, and E. Fosler-Lussier. Structured discriminative models for speech recognition: An overview. *Signal Processing Magazine, IEEE*, 29(6):70 –81, nov. 2012.

[39] H. Hermansky, D.P.W. Ellis, and S. Sharma. Tandem connectionist feature extraction for conventional hmm systems. In *ICASSP*, volume 3, pages 1635 –1638 vol.3, 2000.

[40] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82 –97, nov. 2012.

[41] G. E. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527 –1554, 2006.

[42] G. E. Hinton. A practical guide to training restricted boltzmann machines, 2010.

[43] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors.

[44] Brian Kingsbury, Tara N. Sainath, and Hagen Soltau. Scalable minimum bayes risk training of deep neural network acoustic models using distributed hessian-free optimization. In *INTERSPEECH*, 2012.

[45] Stanley F. Chen. Shrinking exponential language models. In *NAACL*, pages 468–476, 2009.

[46] Holger Schwenk. Continuous space language models. *Computer Speech and Language*, 21(3):492 – 518, 2007.

[47] Ebru Arisoy, Tara N. Sainath, Brian Kingsbury, and Bhuvana Ramadhadran. Deep neural network language models. In *NAACL-HLT*, pages 20–28, 2012.

[48] P. Xu, S. Khudanpur, M. Lehr, E. Prud'hommeaux, N. Glenn, D. Karakos, B. Roark, K. Sagae, M. Saraclar, I. Shafran, D. Bikel, C. Callison-Burch, Y. Cao, K. Hall, E. Hasler, P. Koehn, A. Lopez, M. Post, and D. Riley. Continuous space discriminative language modeling. In *ICASSP*, pages 2129 –2132, march 2012.

[49] T. Oba, T. Hori, A. Nakamura, and A. Ito. Round-robin duel discriminative language models. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(4):1244 –1255, may 2012.

[50] Gakuto Kurata, Abhinav Sethy, Bhuvana Ramabhadran, Ariya Rastrow, Nobuyasu Itoh, and Masafumi Nishimura. Acoustically discriminative language model training with pseudo-hypothesis. *Speech Commun.*, 54(2):219–228, February 2012.

[51] Bjorn Hoffmeister, Tobias Klein, Ralf Schluter, and Hermann Ney. Frame based system combination and a comparison with weighted rover and cnc. In *INTERSPEECH*, pages 537 –540, 2006.

[52] H. Soltau, G. Saon, B. Kingsbury, H.-K.J. Kuo, L. Mangu, D. Povey, and A. Emami. Advances in arabic speech transcription at ibm under the darpa gale program. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(5):884 –894, july 2009.

[53] 河原 達也, 秋田 祐哉, 三村 正人, 政瀧 浩和, and 高橋 敏. 衆議院会議録作成における音声認識システムー全体の構成と評価ー. In **音講論（春）**, 2011.

[54] 小橋川 哲, 浅見 太一, 山口 義和, 阪内 澄宇, 小川 厚徳, 政瀧 浩和, 高橋 敏, and 河原 達也. 衆議院会議録作成における音声認識システムー事前音響処理ー. In **音講論（春）**, 2011.

[55] 三村 正人, 秋田 祐哉, and 河原 達也. 衆議院会議録作成における音声認識システムー音響モデルー. In **音講論（春）**, 2011.

[56] 秋田 祐哉, 河原 達也, and 政瀧 浩和. 衆議院会議録作成における音声認識システムー言語モデルー. In **音講論（春）**, 2011.

[57] 堀 貴明, 中村 篤, 山口 義和, 小橋川 哲, 浅見 太一, 政瀧浩和, 高橋 敏, and 河原 達也. 衆議院会議録作成における音声認識システムー探索技術ー. In **音講論（春）**, 2011.

[58] Nobuaki Minematsu. Yet another acoustic representation of speech sounds. *ICASSP*, 2004.

[59] N. Minematsu, Yu Qiao, S. Asakawa, and M. Suzuki. Speech structure and its application to robust speech processing. *Journal of New Generation Computing*, 28(3):299 –319, 2010.

[60] 鈴木 雅之, 峯松 信明, and 広瀬啓吉. 音声の構造的表象と多段階の重回帰を用いた外国語発音評価. **情報処理学会論文誌**, 52(5):1899 –1909, 2011.

[61] H.G. Hirsch and D. Pearce. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. *Proc. ISCA ITRW ASR*, 2000.

[62] T. Yoshioka and T. Nakatani. Speech enhancement based on log spectral envelope model and harmonicity-derived spectral mask, and its coupling with feature compensation. *Proc. ICASSP*, pages 5064–5067, 2011.

[63] D. Pierce and A. Gunawardana. Aurora 2.0 speech recognition in noise: Update 2. complex backend definition for aurora 2.0, 2002.

[64] Y. Shinohara, T. Masuko, and M. Akamine. Feature enhancement by speaker-normalized splice for robust speech recognition. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4881–4884. IEEE, 2008.

[65] M. Collins and T. Koo. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1):25–70, 2005.

[66] Takanobu Oba, Takaaki Hori, and Atsushi Nakamura. Efficient discriminative training of error corrective models using high-wer competitors. In *Asian Workshop on Speech Science and Technology, IEICE Technical Report SP2007-185-214*, pages 99 –104, 2008.

[67] K. Crammer, R. McDonald, and F. Pereira. Scalable large-margin online learning for structured classification. In *NIPS Workshop on Learning With Structured Outputs*, 2005.

[68] Y. Qiao and N. Minematsu. A study on invariance of f-divergence and its application to speech recognition. *IEEE Transactions on Signal Processing*, 58(7):3884–3890, 2010.

[69] S. Asakawa, N. Minematsu, and K. Hirose. Multi-stream parameterization for structural speech recognition. *ICASSP*, pages 4097–4100, 2008.

[70] E. Arisoy, M. Saraclar, B. Roark, and I. Shafran. Discriminative language modeling with linguistic and statistically derived features. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2):540 –550, feb. 2012.

[71] A. Mohamed, G. Hinton, and G. Penn. Understanding how deep belief networks perform acoustic modelling. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4273 –4276, march 2012.

発表文献

学術論文

[1] M. Suzuki, T. Yoshioka, S. Watanabe, N. Minematsu, K. Hirose, “Feature Enhancement Jointly Using Consecutive Corrupted and Noise Feature Vectors with Discriminative Region Weighting,” IEEE Transaction on Acoustic, Speech and Language Processing (submitted)

[2] 峯松信明，中村新芽，鈴木雅之，平野宏子，中川千恵子，中村則子，田川恭識，広瀬啓吉，橋本浩弥，“日本語アクセント・イントネーションの教育・学習を支援するオンラインインフラストラクチャの構築とその評価”，電子情報通信学会論文誌 (submitted)

[3] 鈴木雅之，黒岩龍，印南圭祐，小林俊平，清水信哉，峯松信明，広瀬啓吉，“条件付き確率場を用いた日本語東京方言のアクセント結合自動推定”，電子情報通信学会論文誌 (2013-3, to appear)

[4] S. Shimizu, M. Suzuki, N. Minematsu, and K. Hirose, “An experimental study on dynamic features of speech structure,” Journal of Research Institute of Signal Processing, vol.16. no.4. pp.319-322 (2012-7)

[5] T. Kai, M. Suzuki, K. Chijiiwa, N. Minematsu, and K. Hirose, “Combination of SPLICE and feature normalization for noise robust speech recognition,” Journal of Research Institute of Signal Processing, vol.16, no.4. pp.323-326 (2012-7)

[6] 峯松信明，鎌田圭，朝川智，鈴木雅之，牧野武彦，西村多寿子，広瀬啓吉，“音声の構造的表象に基づく学習者分類の検証と発音矯正度推定の高精度化”，情報処理学会論文誌，vol.52，no.12，pp.3671-3681 (2011-12)

[7] 鈴木雅之，峯松信明，広瀬啓吉，“音声の構造的表象と多段階の重回帰を用いた外国語発音評価”，情報処理学会論文誌，vol.52，no.5，pp.1899-1909 (2011-5)

[8] 峯松信明，櫻庭京子，西村多寿子，喬宇，朝川智，齋藤大輔，鈴木雅之，“音声に含まれる言語的情報と非言語的情報を分離する手法の提案 〜人間らしい音声情報処理の実現に向けた一検討〜”，電子情報通信学会論文誌，vol.94-D，no.1，pp.12-26 (2011-1)

[9] N. Minematsu, S. Asakawa, Y. Qiao, M. Suzuki, “Speech structure and its application to robust speech processing,” Journal of New Generation Computing (2010-8)

国際会議論文

[10] D. N. Duc, M. Suzuki, N. Minematsu and K. Hirose, “Wideband Re-synthesis of Narrowband Speech Using Discriminative Piecewise Linear Transformation,” Proc. NCSP (2013-3, to appear)

[11] Y. Kashiwagi, M. Suzuki, N. Minematsu, K. Hirose, ”Audio-visual feature integration based on piecewise linear transformation for noise robust automatic speech recognition,” Proc. Spoken Language Technology (SLT), (2012-12)

[12] Y. Luan, M. Suzuki, Y. Yamauchi, N. Minematsu, K. Hirose, ”Performance improvement of automatic pronunciaiton assessment in a noisy classroom,” Proc. Spoken Language Technology (SLT), (2012-12)

[13] T. Zhao, A. Hoshino, M. Suzuki, N. Minematsu, K. Hirose, ”Automatic Chinese pronunciation error detection using SVM with structural features,” Proc. Spoken Language Technology (SLT), (2012-12)

[14] M. Suzuki, G. Kurata, M. Nishimura N. Minematsu, and K. Hirose, “Discriminative reranking for LVCSR leveraging invariant structure,” Proc. INTERSPEECH (2012-9)

[15] 峯松信明，鈴木雅之，平野宏子，中川千恵子，中村則子，田川恭識，広瀬啓吉，”音声出力機能を有したオンライント辞書の構築”，日本語教育国際研究大会予稿集，第一分冊，pp.94 (2012-8)

[16] M. Suzuki, T. Yoshioka, S. Watanabe, N. Minematsu, and K. Hirose, “MFCC enhancement using joint corrupted and noise feature space for highly non-stationary noise environments,” Proc. ICASSP, pp.4109-4112 (2012-3) **Fujitsu student paper award**

[17] K. Chijiiwa, M. Suzuki, N. Minematsu, and K. Hirose, “Unseen noise robust speech recognition using adaptive piecewise linear transformation,” Proc. ICASSP, pp.4289-4292 (2012-3)

[18] S. Shimizu, M. Suzuki, N. Minematsu, and K. Hirose, “An experimental study on dynamic features of speech structure,” Proc. NCSP, pp.166-169 (2012-3) **Student paper award**

[19] T. Kai, M. Suzuki, K. Chijiiwa, N. Minematsu, and K. Hirose, "Combination of SPLICE and feature normalization for noise robust speech recognition," Proc. NCSP, pp.253-256 (2012-3) **Student paper award**

[20] Y. Qiao, M. Suzuki, N. Minematsu and K. Hirose, "Structure constrained distributions matching using quadratic programming and its application to pronunciation evaluation," Proc. ACPR (2011-11)

[21] M. Suzuki, K. Gakuto, M. Nishimura, and N. Minematsu, "Continuous Digits Recognition Leveraging Invariant Structure," Proc. INTERSPEECH, pp.993-996 (2011-9)

[22] S. Kobayashi, S. Shimizu, M. Suzuki, N. Minematsu, K. Hirose, and H. Hirano, "Automatic Generation of Accent Dictionary of Conjugational Words for Any Japanese Texts," Proc. ICJLE, pp.784-786 (2011-8)

[23] M. Suzuki, Y. Qiao, N. Minematsu, and K. Hirose, "Integration of multilayer regression analysis with structure-based pronunciation assessment," Proc. INTERSPEECH, pp.586-589 (2010-9)

[24] M. Suzuki, Y. Qiao, N. Minematsu, and K. Hirose, "Pronunciation proficiency estimation based on multilayer regression analysis using speaker-independent structural fatures," Proc. L2WS, (2010-9) **Best student award**

[25] M. Suzuki, N. Minematsu, D. Luo, and K. Hirose, "Sub-structure-based estimation of pronunciation proficiency and classification of learners," Proc. Int. Workshop on Automatic Speech Recognition and Understanding (ASRU'2009), pp.574-579 (2009-12)

[26] Y. Qiao, M. Suzuki, and N. Minematsu, "A study of Hidden Structure Model and its application of labeling sequences," Proc. Int. Workshop on Automatic Speech Recognition and Understanding (ASRU'2009), pp.118-123 (2009-12)

[27] M. Suzuki, L. Dean, N. Minematsu, K. Hirose, "Improved structure-based automatic estimation of pronunciation proficiency," Proc. ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE), CD-ROM (2009-9)

[28] N. Minematsu and M. Suzuki, "Structure-based pronunciation assessment," Proc. ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE), Demo Session (2009-9)

[29] H. Hirano, M. Suzuki, K. Innami, N. Minematsu, and K. Hirose, “Development of an on-line word accent dictionary of Japanese,” Proc. Int. Conf. on Japanese Language Education (ICJLE’2009) (2009-7)

[30] Y. Qiao, M. Suzuki, and N. Minematsu, “Affine invariant features and its application to speech recognition,” Proc. Int. Conf. Acoustics, Speech, & Signal Processing (ICASSP’2009), pp.4629-4632 (2009-4)

国内研究会・全国大会

[31] 鈴木雅之，峯松信明，広瀬啓吉，コンディション変数の導入によるミスマッチがない場合にも頑健なステレオベース特徴量強調，日本音響学会春季講演論文集，(2013-3，発表予定)

[32] 池島純，鈴木雅之，峯松信明，広瀬啓吉，ターゲット話者の特徴量状態識別に基づく区分的線形変換を用いた声質変換，日本音響学会春季講演論文集，(2013-3，発表予定)

[33] グェンドゥックズイ，鈴木雅之，峯松信明，広瀬啓吉，識別的な区分的線形変換を用いた狭帯域音声に対する帯域拡張，日本音響学会春季講演論文集，(2013-3，発表予定)

[34] 甲斐常伸，鈴木雅之，峯松信明，広瀬啓吉，SPLICE を用いた雑音抑圧手法の統合，日本音響学会春季講演論文集，(2013-3，発表予定)

[35] 槇佑馬，鈴木雅之，橋本浩弥，峯松信明，広瀬啓吉，点予測を用いたアクセント結合自動推定，日本音響学会春季講演論文集，(2013-3，発表予定)

[36] 中村新芽，鈴木雅之，峯松信明，橋本浩弥，中川千恵子，中村則子，平野宏子，田川恭識，広瀬啓吉，日本語教育のための韻律読み上げチュータを備えた Web システムの開発と評価，日本音響学会春季講演論文集，(2013-3，発表予定)

[37] 加藤集平，鈴木雅之，峯松信明，広瀬啓吉，山内豊，西川恵，識別モデルを用いた英語読み上げ文発声の強勢自動評価，日本音響学会春季講演論文集，(2013-3，発表予定)

[38] 橋本浩弥，鈴木雅之，広瀬啓吉，峯松信明，日本語 HMM 音声合成におけるコンテキストラベルの改良，日本音響学会春季講演論文集，(2013-3，発表予定)

[39] 岡安貴大，池島純，柏木陽佑，鈴木雅之，峯松信明，広瀬啓吉，実環境下における GMM を用いた統計的声質変換の検討，日本音響学会春季講演論文集，(2013-3，発表予定)

[40] 加藤集平，鈴木雅之，峯松信明，広瀬啓吉，山内豊，西川惠，識別モデルを用いた英文読み上げ音声からの強勢自動検出，電子情報通信学会音声研究会資料，(2012-2，発表予定)

[41] 鈴木雅之，倉田岳人，西村雅史，峯松信明，広瀬啓吉，音声の構造的表象を用いた大語彙音声認識の識別的リランキング，日本音響学会秋季講演論文集，2-1-4，pp.63-66 (2012-9) **栗屋潔学術奨励賞**

[42] 鈴木雅之，黒岩龍，印南圭祐，小林俊平，清水信哉，峯松信明，広瀬啓吉，CRF を用いた日本語東京方言のアクセント結合自動推定，日本音響学会秋季講演論文集，2-2-12，pp.299-302 (2012-9)

[43] 柏木陽佑，鈴木雅之，峯松信明，広瀬啓吉，区分的線形変換を用いた雑音環境下マルチモーダル音声認識，日本音響学会秋季講演論文集，1-1-9, pp.25-28 (2012-9)

[44] 甲斐常伸，鈴木雅之，峯松信明，広瀬啓吉，雑音抑圧・特徴量強調・特徴量正規化を組み合わせた雑音に頑健な大語彙音声認識，日本音響学会秋季講演論文集，1-1-8，pp.21-24 (2012-9)

[45] 正木大介，鈴木雅之，峯松信明，広瀬啓吉，雑音環境下音声認識のための長時間セグメント特徴量に関する検討，日本音響学会秋季講演論文集，1-1-10，pp.29-32 (2012-9)

[46] ランイ，鈴木雅之，加藤集平，峯松信明，広瀬啓吉，Performance improvement of automatic pronunciation assessment in noisy classroom，日本音響学会秋季講演論文集，3-2-7，pp.45-46 (2012-9)

[47] Teiseki Ou, Masayuki Suzuki, Nobuaki Minematsu, Kyoko Sakuraba, and Keikich Hirose, GMM-スーパーベクトルと SVM に基づく GID 話者ための女性度推定，日本音響学会秋季講演論文集，1-1-15，pp.45-46 (2012-9)

[48] Tongmu Zhao, Hoshino Akemi, Masayuki Suzuki, Minematsu Nobuaki, and Keikichi Hirose, SVM と構造表象に基づく中国語発音誤りの自動検出，日本音響学会秋季講演論文集，3-Q-25，pp.415-418 (2012-9)

[49] Tongmu Zhao, Masayuki Suzuki, Minematsu Nobuaki, and Keikichi Hirose, Automatic pronunciation error detection of Chinese using SVM with structural features, IEICE Technical Report，SP (2012-7)

[50] 柏木陽佑，鈴木雅之，峯松信明，広瀬啓吉，SPLICE に基づく音声・口唇画像情報を用いた雑音環境下音声認識，電子情報通信学会音声研究会資料，SP-2012-27，pp.155-160 (2012-5)

[51] 甲斐常伸，鈴木雅之，峯松信明，広瀬啓吉，雑音抑圧と SPLICE を組み合わせた雑音環境下音声認識，電子情報通信学会音声研究会資料，SP-2012-28，pp.161-166 (2012-5)

[52] 鈴木雅之，吉岡拓也，渡部晋司，峯松信明，広瀬啓吉，クリーン音声状態の識別に基づく特徴量強調，日本音響学会春季講演論文集，1-7-10，pp.23-26 (2012-3)

[53] 千々岩圭吾，鈴木雅之，峯松信明，広瀬啓吉，主成分分析を用いた GMM に基づく耐雑音音声認識フロントエンドの高精度化，日本音響学会春季講演論文集，1-P-17，pp.161-164 (2012-3)

[54] 清水信哉，鈴木雅之，峯松信明，広瀬啓吉，トラジェクトリモデルを用いた音声の構造的表象，日本音響学会春季講演論文集，1-P-3，pp.123-126 (2012-3)

[55] 甲斐常伸，鈴木雅之，峯松信明，広瀬啓吉，特徴量正規化と SPLICE を組み合わせた雑音環境下音声認識，日本音響学会春季講演論文集，1-7-5，pp.37-40 (2012-3)

[56] 加藤集平，鈴木雅之，峯松信明，広瀬啓吉，GOP と回帰分析を用いたシャドーイング評価の高精度化，日本音響学会春季講演論文集，3-11-18，pp.417-420 (2012-3)

[57] 黒瀧夏子，鈴木雅之，峯松信明，広瀬啓吉，構造的表象を用いた話者間の発音距離行列の可視化に関する検討，日本音響学会春季講演論文集，1-R-25，pp.491-494 (2012-3)

[58] Nguyen Duc Duy，鈴木雅之，齋藤大輔，峯松信明，広瀬啓吉，構造表象を用いた音声認識における状態数決定に関する実験的検討，日本音響学会春季講演論文集，1-P-6，pp.131-132 (2012-3)

[59] 鈴木雅之，倉田岳人，西村雅史，峯松信明，音声の構造的表象を用いた連続数字音声認識，日本音響学会秋季講演論文集，1-10-14 (2011-9) **学生奨励賞**

[60] 鈴木雅之，吉岡拓也，峯松信明，広瀬啓吉，非定常雑音環境における線形判別分析を用いた静的・動的 MFCC の強調，日本音響学会秋季講演論文集，1-10-6 (2011-9)

[61] 清水信哉，鈴木雅之，峯松信明，広瀬圭吉，音声の構造的特徴の動的特徴，日本音響学会秋季講演論文集，1-10-2，(2011-9)

[62] 千々岩圭吾，鈴木雅之，峯松信明，広瀬啓吉，Eigen-SPLICE を用いた雑音環境下における音声認識の実験的検討，日本音響学会秋季講演論文集，2-Q-19 (2011-9) **学生奨励賞**

[63] T. Zhao, M. Suzuki, C. Wang, N. Minematsu, K. Hirose, Chinese language pronunciation proficiency estimation based on structural features, 日本音響学会秋季講演論文集，3-Q-24 (2011-9)

[64] 千々岩圭吾, 鈴木雅之, 齋藤大輔, 峯松信明, 広瀬啓吉，Eigen-SPLICE を用いた雑音環境下における音声認識，情報処理学会音声言語情報処理研究会，2011-SLP-87-15 (2011-7)

[65] 鈴木雅之，吉岡拓也，渡部晋治，峯松信明，雑音特徴量を用いた劣化音声特徴量変換に関する検討，日本音響学会春季講演論文集，1-Q-15，pp.795-798 (2011-3)

[66] 甲斐常伸，鈴木雅之，齋藤大輔，峯松信明，広瀬啓吉，孤立音を対象にした構造的表象の理論的考察と実験的検討，日本音響学会春季講演論文集，2-P-2，pp.139-142 (2011-3)

[67] 鈴木雅之，峯松信明，広瀬啓吉，構造的特徴量を用いた CALL システムの開発，日本音響学会秋季講演論文集，3-P-27，pp.393-394 (2010-9)

[68] 鈴木雅之，中村綾乃，喬宇，峯松信明，広瀬啓吉，構造的特徴量に対する多段階の重回帰分析による発音評価，電子情報通信学会音声研究会，SP2010-37，pp.13-18 (2010-7)

[69] 清水信哉，齋藤大輔，鈴木雅之，峯松信明，広瀬啓吉，用法の違いを考慮した類似単語の置換とそれを用いた言語モデル学習データ自動生成，人工知能学会全国大会講演集，2G1-OS3-4，pp.1-4 (2010-6)

[70] 清水信哉，鈴木雅之，齋藤大輔，峯松信明，広瀬啓吉，用法の違いを考慮した類似単語の置換による学習データ生成とそれを用いた主題の違いに頑健な言語モデルの構築，情報処理学会音声言語情報処理研究会，2010-SLP-81-9 (2010-5) **学生奨励賞**

[71] 鈴木雅之，喬宇，峯松信明，広瀬啓吉，牧野武彦，“構造表象と多段階の重回帰を用いた外国語発音評価”，日本音響学会春季講演論文集 (2010-3)

[72] 中村綾乃，鈴木雅之，峯松信明，広瀬啓吉，牧野武彦，“音声の構造的表象を用いた英語二重母音の発音評価に関する検討”，日本音響学会春季講演論文集 (2010-3)

[73] 千々岩圭吾，鈴木雅之，齋藤大輔，峯松信明，広瀬啓吉，“非周期性に注目した基本周期パターン生成過程モデルのパラメータ自動抽出の高精度化”，日本音響学会春季講演論文集 (2010-3)

[74] 清水信哉，齋藤大輔，鈴木雅之，峯松信明，広瀬啓吉，“類似単語の置換による言語モデルの平滑化”，日本音響学会春季講演論文集 (2010-3)

[75] 鈴木雅之，喬宇，峯松信明，広瀬啓吉，“音声の構造的表象と多段階の重回帰を用いた外国語発音分析”，情報処理学会全国大会講演集，1U-9 (2010-3) **学生奨励賞**

[76] 鈴木雅之，羅徳安，峯松信明，広瀬啓吉，“発音構造を用いた話者の違いに頑健な発音評定と学習者分類”，日本音響学会秋季講演論文集，1-2-5，pp.243-246 (2009-9)

[77] 喬宇，鈴木雅之，峯松信明，“Proposal of Hidden Structure Model,” 日本音響学会秋季講演論文集，1-1-3，pp.7-10 (2009-9)

[78] 鈴木雅之，羅徳安，峯松信明，広瀬啓吉，“音声の構造的表象を用いた自動発音評定法の改善”，情報処理学会音声言語情報処理研究会，2009-SLP-77-17，pp.1-6 (2009-7)

[79] Y. Qiao, M. Suzuki, and N. Minematsu, “An Investigation of Hiden Structure Model,” 情報処理学会音声言語情報処理研究会, 2009-SLP-77-5, pp.1-6 (2009-7)

[80] 鈴木雅之，喬宇，峯松信明，広瀬啓吉，“アフィン変換不変性を有する局所特徴量を用いた音声認識”，日本音響学会春季講演論文集，1-5-4，pp.11-14 (2009-3)

[81] 國越晶，喬宇，鈴木雅之，峯松信明，広瀬啓吉，坂野秀樹，“ジェスチャー空間と音響空間の写像に基づくリアルタイム音声生成系”，日本音響学会春季講演論文集，2-P-2，pp.445-448 (2009-3)

[82] 鈴木雅之，喬宇，峯松信明，広瀬啓吉，“アフィン変換不変性を有する局所的特徴量を用いた音声認識”，電子情報通信学会音声研究会，SP2008-114，pp.209-214 (2008-12)

[83] 國越晶，喬宇，鈴木雅之，峯松信明，広瀬啓吉，“空間写像に基づく手の動きを入力とした音声生成系の構築”，電子情報通信学会音声研究会，SP2008-78，pp.45-50 (2008-11)

[84] 鈴木雅之，朝川智，喬宇，峯松信明，広瀬啓吉，“スペクトル領域特徴量を用いた音声の構造的表象による音声認識”，日本音響学会秋季講演論文集，1-R-26，pp.473-476 (2008-9)

[85] 鈴木雅之，朝川智，喬宇，峯松信明，広瀬啓吉，“スペクトル特徴量を用いた音声の構造的表象に関する実験的検討”，電子情報通信学会音声研究会，SP2008-32，pp.73-78 (2008-6)

[86] 桜庭京子，峯松信明，広瀬啓吉，鈴木雅之，田山二朗，今泉敏，山内俊雄，“MtF のボイスセラピーにおける成功症例の型の分類”，性同一性障害学会研究大会 (2008-3)